JCTC Journal of Chemical Theory and Computation

On Ion and Molecular Polarization of Halides in Water

Elvira Guàrdia and Ioannis Skarmoutsos

Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord B4-B5, Barcelona 08034, Spain

Marco Masia*

Dipartimento di Chimica, Università degli Studi di Sassari, Sardinian Laboratory for Computational Materials Science SLACS (INFM-CNR) and INSTM, Via Vienna 2, 07100 Sassari, Italy

Received February 26, 2009

Abstract: The high polarizability of halide anions affects, in aqueous solutions, many phenomena ranging from hydrogen bond dynamics to water interfaces' structure. In this Letter dipolar interactions of halides in water are investigated through Car−Parrinello Molecular Dynamics simulations. Contrary to previous studies, a different polarization of first and second hydration shell water molecules is found. The analysis hints that existing classical polarizable force fields lack a description of short-range interactions which causes an overestimation of polarization effects.

I. Introduction

It is widely accepted that molecular polarization affects many properties shown by inhomogeneous systems. In particular, in ionic solutions, although dipolar interactions decay faster than coulomb interactions, they are considered to be responsible for macroscopic properties in the bulk phase and at interfaces.[1−13]

In spite of its importance, there are few simulation studies available where the features of polarizable interaction are studied in detail.[14−20] This task is currently pursued in our group. To this end, reference data to compare with are needed. In particular ab initio calculations are the only source for defining the electrostatic properties of halide−water solutions. Recently, ab initio MD simulations[21] of halide anions dissolved in water have been performed by a few authors.[2,22−27] Most of them made use of Car−Parrinello Molecular Dynamics for the whole systems, except for ref 26, where the ion was described quantum chemically with a self-consistent field model, while the solvent molecules were described classically (including many body effects). A mixed DFT/MM Monte Carlo approach for studying the bromide ion in water has also been used by Tuñon et al.[28] The above studies focus mainly on the structure of the first solvation shell and on its dynamics. Electrostatic properties are usually not considered or they are just mentioned; to our knowledge there is not a systematic comparison of electrostatic properties of halide anions in water solution. We are aware only of the study of Krekeler et al.,[29] who performed first principle density functional calculations to look into the properties of small $[X(H_2O)_n]^-$ clusters ($n = 1,2,...,6$), X being fluoride, chloride or iodide. They found that, as the number of water molecules increases, molecular polarization is determined by water−water interactions rather than by ion−water interactions; in fact, in their calculations the dipole moments of first shell molecules tend to the same value of bulk water. This conclusion led them to support the use of nonpolarizable classical force fields, contrary to what is suggested by many authors. While, on the one hand, it is true that the dipole moment of water molecules in the solvation shells is closer to that of bulk water rather than to that of gas phase clusters with small n, on the other hand, it should be considered that the hydration shell is a dynamical entity, which changes in time causing the instantaneous induced dipoles to be much different than in the bulk. It would be impossible to model such dynamic response to the change in the solvation environment by using simple nonpolarizable force fields (*vide infra*). Furthermore, to study the influence of the ion on all solvation shells, larger systems should be considered. In the present work, in order to have an insight on polarizable interactions beyond the second solvation shell, we have carried out a study of electrostatic properties of halide anions dissolved in 96 water molecules by employing Car−Parrinello MD simulations. These simulations are meant to be a reference for future comparisons with polarizable classical force fields. In this contribution the contents are organized as follows: in the next section we present the details of our calculations; then, the electrostatic properties of the ion and of the water molecules belonging to different solvation shells are discussed in section III. Finally our conclusions are briefly summarized in the last section.

II. Computational Details

Ab initio MD simulations were performed using the Car−Parrinello (CP)[30] scheme for propagating the wave functions and the ionic configurations as implemented in the CPMD package.[31] In the present study we have used dispersion-corrected atom-centered pseudopotentials (DCACPs)[32,33] in

* Corresponding author e-mail: marco.masia@uniss.it.

10.1021/ct900096n CCC: $40.75
Published on Web 05/18/2009

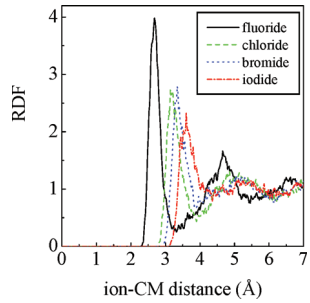


Figure 1. Radial distribution function between the ion and the water center of mass.

the Troullier−Martins[35] format for oxygen and hydrogen. It has been recently shown that DCACPs are successful in accounting for London dispersion forces and that they are capable of faithfully reproducing many dynamical and structural properties of water.[33] For halides, since these kinds of pseudopotentials are still under development,[34] we have used norm-conserving Goedecker pseudopotentials.[36−38] The BLYP density functional[39,40] was used for the electronic strycture calculations. The cutoff for the wave function was set to 80 Ry, the time step was set to 4 au, and the fictitious mass for the orbital was chosen to be 400 amu The length of the cubic simulation box was computed in order to get a total density of 1 g cm^{-3}. Periodic boundary conditions were applied. Production runs of 15 ps in the microcanonical ensemble followed NVT equilibration runs of 3 ps where the temperature was set to 330 K; the initial configurations were generated with classical molecular dynamics simulations of 200 ps. The same procedure has been applied to simulate pure water (96 molecules).

Every five time steps, the Wannier centers'[41,42] coordinates $\mathbf{r}_j(t)$ were computed; in this way, given the ionic coordinates $\mathbf{R}_i(t)$, the dipole moment of a molecule (or of the anion) I were obtained as

$$\boldsymbol{\mu}_I = \sum_{i=1}^{N_I} Q_i \mathbf{R}_i + \sum_{j=1}^{n_I} q_j \mathbf{r}_j \qquad (1)$$

with Q_i and q_j being respectively the charge of the i^{th} ion and of the j^{th} Wannier center.

III. Results and Discussion

Radial distribution functions between the anions and the water center of mass are shown in Figure 1. First and second hydration shells can be clearly devised for all ions. It could be appreciated that the radius of the first shell increases in the group (R_{F^-} = 3.32 Å, R_{Cl^-} = 3.90 Å, R_{Br^-} = 3.90 Å, R_{I^-} = 4.36 Å) as well as the hydration number (N_{F^-} = 4.9, N_{Cl^-} = 6.5, N_{Br^-} = 6.5, N_{I^-} = 8.5), in agreement with previous results. A deeper study of the structural properties and of water dynamics in these systems will be presented in a short coming full paper. Here we focus on the electrostatic properties within each hydration shell.

In Table I the average values of the ion and water dipole moments are shown. It can be clearly seen that, as the ion polarizability[43] increases, the average dipole moment increases as well. This is what was expected and what was found with previous ab initio calculations at condensed phase. The calculations of Öhrn and Karlström[26] give the same value of ours for chloride, while, for fluoride, the induced dipole moment is the double of what we get. For bromide, Tuñon et al.,[28] with DFT/MM Monte Carlo calculations, found a value five times lower than ours. We believe that the disagreement between our results and the above cited studies is due to the treatment of the solvent with classical models which could not faithfully describe the polarizable feedback between water molecules and the anion. This is supported by the comparison with previous Car−Parrinello MD simulations,[22−25] where also the solvent is treated at the same quantum chemical level as the anion; in Table I it can be seen that the average values for the anion are similar to previous ab initio simulations (with a maximum difference of ca. 0.1 D). In our calculations we get a smaller standard deviation (which was calculated with the proper methods for correlated data sets);[44] it is probably due to the fact that we performed longer simulations (gathering much more configurations to average over) and that the system size is larger, which causes the amplitude of fluctuations to be lower.

Table I. Average Dipole Moments and Their Standard Deviations for the Ion and First and Second Shell Molecules[a]

	ion $\langle\mu\rangle$ (σ_μ)	first shell $\langle\mu\rangle$ (σ_μ)	second shell $\langle\mu\rangle$ (σ_μ)
water-F^-	0.42 (0.18)	3.04 (0.31)	2.96 (0.30)
ref 25	0.39	3.07 (0.30)	3.10 (0.30)
ref 26	0.19 (0.06)	--	--
water-Cl^-	0.82 (0.32)	2.87 (0.27)	2.95 (0.29)
ref 23	1.00	3.14 (0.57)	3.15 (0.65)
ref 26	0.89 (0.36)	--	--
ref 27	--	3.07	3.07
water-Br^-	1.02 (0.40)	2.87 (0.27)	2.98 (0.29)
ref 22	0.9 (0.8)	2.9	2.9
ref 28	0.21	--	--
water-I^-	1.21 (0.51)	2.83 (0.27)	2.92 (0.29)
ref 24	1.3 (1.1)	3.0 (0.6)	3.0 (0.6)

[a] Units: Debye.

If the dipole moments of the first and second hydration shells are considered, our data are still in fair agreement with previous ab initio simulations, the bigger differences being due to the different treatment of the electronic structure calculations in our simulations (see discussion on DCACPs in section II). In passing we should mention that the dipole moment obtained for pure water is 2.96 ± 0.30 D, in agreement with previous ab initio simulations of Silvestrelli and Parrinello[42] (who obtained μ = 2.95 D).[45] Even if all the values for the solvation shell water molecules are within the error, it should be pointed out that, in our calculations, there are always differences of ca. 0.1 D between the average dipole moments of the first and of the second solvation shell (the latter being very close to the bulk water dipole moment). Such a trend was not observed in previous simulations, probably because of the lower accuracy in the statistics (see above). In this aspect our work is the first where a different polarization of first and second hydration shell molecules in water−halide solutions has been found. Although such a difference in the water polarization is not that big if related to the total dipole moment (only a 3%), it should be noted that it constitutes approximately 10% of the induced dipole moment, which is due to the balance of the electrostatic interactions with the ion and with other water molecules. The difference in the induced dipole moment might be due to the fact that the negative charge on the ion is somewhat screened

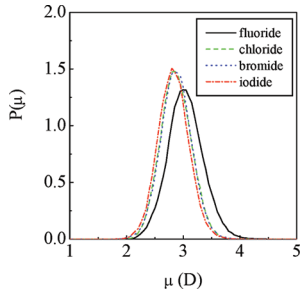


Figure 2. Probability distribution function for the dipole moment of water molecules in the first hydration shell.

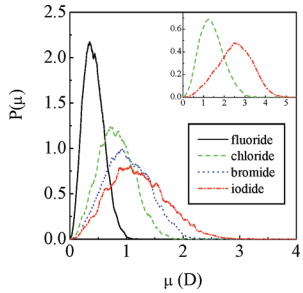


Figure 3. Probability distribution function for the dipole moment of halide anions. In the inset the PDFs of classical simulations of chloride (dots) and iodide (dash−dots) in water are shown; the latter, kindly provided by Wick and Xantheas, appears in Figure 6 of ref 49.

by the first shell molecules, causing second shell molecules to be less polarized. Similar results were recently obtained with Car−Parrinello MD simulations of water−cation solutions (the cations being K^+ and Ca^{2+}).[48]

A feature which particularly strikes one's attention is the different polarization of first shell molecules in the fluoride−water system. While in the case of chloride, bromide, and iodide the average dipole moment of first shell water molecules is ca. 2.85 D, i.e. 0.1 D lower than bulk water, in the case of fluoride, the dipole moment is ca. 0.2 D higher. The difference could be better appreciated from an inspection of the probability distribution functions (PDFs) for the first shell molecules in each system considered (Figure 2). The different response of water molecules in the first solvation shell of fluoride was previously observed in small water−halide clusters with Car−Parrinello MD simulations (see Figure 3(c) of ref 29); in fact for a small number of water molecules, the induced dipole moment is higher than for other halides. Since the values converge for 4 and 5 water molecules the authors concluded that, adding more molecules, would only cause the dipole moment to be the same as in the bulk. It should be considered that the first hydration shell of halide anions contains more than five water molecules; moreover, for each cluster, only one configuration was studied. Since the solvation shell is a dynamical entity, where both the structure and the hydration number fluctuate, we believe that the above clusters cannot be considered as representative samples of the first solvation shell. The difference in the behavior of the fluoride solution with respect to the solution of the other anions could be ascribed to the low polarizability of F^-. To better understand this point it is useful to compare the PDFs of the ionic dipole moments as shown in Figure 3. First of all it can be noticed that the PDFs extend for more than 3 D for the most polarizable ions; on the other hand, the fluoride's PDF tail is slightly higher than 1 D. If one considers that fluoride is not as polarized (the induced dipole moment is small) as the rest of the halides in the series, it seems reasonable to expect the dipole−dipole interaction between the ion and water to be small; in this case the coulomb repulsion between the negatively charged ion and the water electron cloud dominates, causing the induced dipole moment on water to be higher than in the bulk. Following the same reasoning, when we consider bigger anions, as the dipole−dipole interaction becomes stronger, the coulomb repulsion does not longer dominate; overall, it all brings to a decrease of the induced dipole moment on water; such an effect is slightly more evident in iodide−water than in chloride−or bromide−water. On top of that, it should be noticed that the damping effect is stronger for bigger anions;[15] it means that the strength of charge−charge, charge−dipole, and dipole−dipole interactions does not increase monotonically as the ion approaches the water molecule but rather shows a turnover at short distances (within the range of the first shell radius).[15] The damping effect is more evident if our results for chloride− and iodide−water solutions are compared with the recent results of the same systems studied with classical molecular dynamics.[49] Until recently, there has been high uncertainty on the values of halide dipole polarizability. This was basically due to the inaccuracy of experimental data on one side and to the low level of ab initio calculations on the other side. According to accurate high level quantum chemical calculation,[50] the dipole polarizability values are 2.47, 5.48, 7.27, and 10.27 $Å^3$ for F^-, Cl^-, Br^-, and I^-, respectively. These values are higher than the ones[51] taken for granted until then (respectively 1.38, 3.94, 5.22, and 7.81 $Å^3$); however, nowadays it is still common to find classical MD or Monte Carlo studies of anions where the old values are considered. It is known, in fact, that using high anion polarizabilities, the dipole moment increases up to values which do not have physical sense. The common approach, then, is to use low polarizabilities rather than to damp the polarizable interactions. Nevertheless, as already stated in refs 15 and 16 such practice only serves to hinder hyperpolarization but not to reproduce the response of the ion at short anion−water distances. In the inset of Figure 3 we show the PDFs as obtained in classical MD simulations[49] using $\alpha_{Cl^-} = 4.5$ $Å^3$ and $\alpha_{I^-} = 6.9$ $Å^3$; the ionic dipole moments are peaked at ca. 1.3 and 2.5 D, i.e. where the respective ab initio PDFs are already decreasing to zero. This big difference among classical and ab initio results should be taken into account when the properties of simulated systems depend on the ionic and molecular polarization. For example, the high propensity for surface states of halide anions in water is explained through the anion induced dipole moments as obtained with classical simulations.[8] Doubts about the physical sense of using gas phase polarizabilities in condensed phase simulations could arise if one takes into account the common view according to which, in the liquid, there is a reducing effect due to electron clouds interactions. Nonetheless, in contrast to this view, it has been recently demonstrated by means of Car−Parrinello MD simulations that both the polarizability tensor of water molecules and of fluoride in CsF at liquid state are distributed around the gas phase values with a narrow distribution.[52] This result encourages the use of gas phase polarizabilities in classical MD simulations,

which, additionally, allows for a higher transferability of the force field from gas to condensed phase simulations.

IV. Concluding Remarks

Although accurate polarizable force fields for modeling (halide) ion−water interactions are needed, up to now few studies have pointed out the need for benchmark data on the electrostatic properties of such systems. We have performed Car−Parrinello MD simulations of one halide ion with 96 water molecules. Such a big system is needed to account for the long-range nature of electrostatic interactions which spread up to the bulk. In previous simulations, the systems were barely formed by the first and second hydration shell of the ion, which would not allow the studying of the differences in the polarizable interaction between those shells and the bulk. To our knowledge this is the first study where the (expected) differences in the polarization of first and second solvation shells could be devised. In previous studies those differences could not be found given the small size of the systems and the short simulated times. Chloride, bromide, and iodide, as an average trend, do polarize first shell water molecules such that their molecular dipole moment is lower than the one of second shell (and bulk) molecules, the difference being ca. 0.1 D (which represents approximately a 10% of the induced dipole moment). Fluoride, on the contrary, tends to overpolarize first shell water molecules so that their molecular dipole moment is higher by ca. 0.1 D than the one of second shell molecules.

The broadness of molecular and ionic dipole moment PDFs justifies the implementation of polarizable models in classical simulations which could account for dynamical effects due to dipolar interactions. The implementation of new ion−water polarizable models cannot be made without considering *Pauli effects* which extend beyond the first solvation shell. Existing classical force fields tend to include those effects by simply reducing the ionic polarizability. We have shown that this is not enough as the induced dipole moment in our ab initio simulations is more than 1 D smaller than what was found in classical simulations.

Acknowledgment. The authors gratefully acknowledge Dr. Collin Wick and Dr. Sotiris Xantheas for having provided their data from classical simulations appearing in the inset of Figure 3 and for the useful discussion about the inclusion of damping effects in classical force fields. We would like to thank Dr. I.-Chun Lin for useful discussions and suggestions about DCACPs. We thankfully acknowledge the computer resources, technical expertise, and assistance provided by the Barcelona Supercomputing Center - Centro Nacional de Supercomputación for the project QCM-2008-2-0010. I.S. acknowledges the postdoctoral financial support from the Department of Physics and Nuclear Engineering of the Technical University of Catalonia. E.G. acknowledges financial support from the Direcció General de Recerca de la Generalitat de Catalunya (Grant 2005SGR-00779) and from the Ministerio de Educación y Ciencia of Spain (Grant FIS2006-12436-C02-01). The research institution INSTM is acknowledged by M.M., who is also thankful for the resources given by the *Cybersar Project* managed by the "Consorzio COSMOLAB".

References

(1) *J. Chem. Theory Comput*. has recently devoted a special issue (2007, vol. 3 p 1877) on polarization highlighting that new and sophisticated polarizable force fields are required to provide more quantitative and converged results.

(2) Mallik, B. S.; Semparithi, A.; Chandra, A. *J. Chem. Phys.* **2008**, *129*, 194512.

(3) Laage, D.; Hynes, J. T. *Proc. Natl. Acad. Sci U. S. A.* **2007**, *104*, 11167.

(4) Smith, J. D.; Saykally, R. J.; Geissler, P. L. *J. Am. Chem. Soc.* **2007**, *129*, 13847.

(5) Omta, A. W.; Kropman, M. F.; Woutersen, S.; Bakker, H. J. *Science* **2003**, *301*, 347.

(6) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. *J. Phys. Chem. B* **2007**, *111*, 13570.

(7) Lo Nostro, P.; Ninham, B. W.; Milani, S.; Lo Nostro, A.; Pesavento, G.; Baglioni, P. *Biophys. Chem.* **2006**, *124*, 208.

(8) Jungwirth, P.; Tobias, D. J. *Chem. Rev.* **2006**, *106*, 1259.

(9) Kuo, I-F. W.; Mundy, C. J. *Science* **2004**, *303*, 658.

(10) Cacace, M. G.; Landau, E. M.; Ramsden, J. J. *Rev. Biophys.* **1997**, *30*, 241.

(11) Coudert, F.-X.; Vuilleumier, R.; Boutin, A. *ChemPhysChem* **2006**, *7*, 2464.

(12) Craig, V. S. J.; Ninham, B. W.; Pashley, R. M. *Nature* **1993**, *364*, 317.

(13) Lugli, F.; Zerbetto, F. *ChemPhysChem* **2007**, *8*, 47.

(14) Masia, M.; Probst, M.; Rey, R. *J. Chem. Phys.* **2005**, *123*, 164505.

(15) Masia, M.; Probst, M.; Rey, R. *Chem. Phys. Lett.* **2006**, *420*, 267.

(16) Masia, M. *J. Chem. Phys.* **2008**, *128*, 184107.

(17) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.

(18) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2005**, *123*, 164108.

(19) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, *111*, 8170.

(20) Söderhjelm, P.; Öhrn, A.; Ryde, U.; Karlström, G. *J. Chem. Phys.* **2008**, *128*, 014102.

(21) It is still debated if DFT methods should be considered *ab initio*, considering that the development of functionals requires the use of parameters. Since the number of parameters is low (compared to other semiempirical methods), and since the quality of results is comparable to wavefunction based methods, we prefer to classify DFT as an ab initio technique.

(22) Raugei, S.; Klein, M. *J. Chem. Phys.* **2002**, *116*, 196.

(23) Heuft, J. M.; Meijer, E. J. *J. Chem. Phys.* **2002**, *119*, 11788.

(24) Heuft, J. M.; Meijer, E. J. *J. Chem. Phys.* **2005**, *122*, 094501.

(25) Heuft, J. M.; Meijer, E. J. *J. Chem. Phys.* **2005**, *123*, 094506.

(26) Öhrn, A.; Karlström, G. *J. Phys. Chem. B* **2004**, *108*, 8452.

(27) Petit, L.; Vuilleumier, R.; Maldivi, P.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 1040.

(28) Tuñon, I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-López, M. F. *Chem. Phys. Lett.* **1995**, *241*, 450.

(29) Krekeler, C.; Hess, B.; Delle Site, L. *J. Chem. Phys.* **2006**, *125*, 054305.

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

(31) CPMD version 3.11, Copyright IBM Corp. 1990−2006, MPI für Festköorperforschung Stuttgart 1997−2001. For downloads see http://www.cpmd.org (accessed Apr 1, 2009).

(32) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(33) Lin, I.-C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. *J. Phys. Chem. B* **2009**, *13*, 1127.

(34) Dr. I.-Chun Lin, private communication.

(35) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993.

(36) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703.

(37) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641.

(38) Krack, M. *Theor. Chem. Acc.* **2005**, *114*, 145.

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

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

(41) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847.

(42) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308. erratum: Phys. Rev. Lett. 1999, 82, 5415.

(43) It is well-known that halide polarizabilities increase in the group as $\alpha_{F^-} < \alpha_{Cl^-} < \alpha_{Br^-} < \alpha_{I^-}$; for the most recent estimates of their value, check ref 50.

(44) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, USA, 2002.

(45) This value is also in good agreement with recent accurate experimental results (see ref 46) according to which $\mu = 2.9 \pm 0.6$ D. Though, on the one hand it should be noted that a direct measurement of the dipole moment of liquid water does not exist, the best estimate being based on the analysis of the X-ray structure; on the other hand, the method employed here allows to compute the *static* dipole moment neglecting the *dynamical* contributions. For a thorough explanation of this issue we refer to ref 47.

(46) Badyal, Y. S.; Saboungi, M.-L.; Price, D. L.; Shastri, S. D.; Haeffner, D. R.; Soper, A. K. *J. Chem. Phys.* **2000**, *112*, 9206.

(47) Pasquarello, A.; Resta, R. *Phys. Rev. B* **2003**, *68*, 174302.

(48) Bucher, D.; Kuyucak, S. *J. Phys. Chem. B* **2008**, *112*, 10786.

(49) Wick, C. D.; Xantheas, S. S. *J. Phys. Chem. B* **2009**, *113*, 41414146.

(50) Hättig, C.; Hess, B. A. *J. Chem. Phys.* **1998**, *108*, 3863.

(51) Coher, J. *J. Phys. Chem.* **1976**, *80*, 2078.

(52) Salanne, M.; Vuilleumier, R.; Madden, P. A.; Simon, C.; Turq, P.; Guillot, B. *J. Phys.: Condens. Matter* **2008**, *20*, 494207.

CT900096N

JCTC Journal of Chemical Theory and Computation

Non-Hermitian Multiconfiguration Molecular Mechanics

Oksana Tishchenko* and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received February 13, 2009

Abstract: We present a new version of the multiconfiguration molecular mechanics (MCMM) algorithm for fitting potential energy surfaces of complex reactive systems. The main improvement consists in allowing the valence bond configuration interaction matrix to be non-Hermitian, which broadens the range of geometries over which the potential energy surface can be fit accurately. A second improvement is that the new algorithm has simpler gradients and Hessians and executes faster. The performance of the new algorithm is evaluated using the example of two model reactions.

1. Introduction

The representation of potential energy surfaces for chemical reaction dynamics continues to present a multifaceted challenge. For small systems, there has been great progress with new approaches for fitting surfaces.[1−7] For larger systems, the only practical approach is often direct dynamics; in this approach, instead of using a prefitted potential energy surface, "all required energies and forces for each geometry that is required for evaluating dynamical properties are obtained directly from electronic structure calculations."[8] This raises the cost unless inexpensive electronic structure methods (such as ab initio Hartree−Fock theory,[9,10] neglect-of-differential-overlap molecular orbital theory,[8,11] or diatomics-in-molecules valence bond theory[12]) are used. However, more reliable results can be obtained if direct dynamics calculations are based on density functional theory,[13,14] multiconfiguration ab initio wave function theory,[15] or multicoefficient correlation methods.[16] Therefore, a variety of approaches intermediate between straight direct dynamics and straight fitting have arisen, such as use of specific reaction parameters[17,18] and low-dimensionality interpolatory methods.[19,20] In the same spirit is the use of combined quantum mechanical and molecular mechanical methods[21] or methods that combine semiempirical valence bond theory[22] with molecular mechanics valence bond diabatic states, as is done in modeling ionic−covalent interactions,[23] in combining valence bond theory for reactive atoms with molecular mechanics for spectator atoms,[24] in the empirical valence bond method of Warshel and Weiss,[25] or in the approximate valence bond method of Bala et al.[26] One very promising method of the latter type is multiconfiguration molecular mechanics[27−34] (MCMM), which also combines semiempirical valence bond theory with molecular mechanics valence bond diabatic states, but in a way that in principle allows systematic improvement of potential energy surfaces to an arbitrarily high accuracy.

Multiconfiguration molecular mechanics was shown[27,28,30,31] to be successful for gas-phase kinetics when the dynamical calculations are based on variational transition state theory[35−37] with multidimensional tunneling[8,38−42] even with large-curvature tunneling approximations[8,39,42] that require energies at points significantly removed from the minimum energy path, and for liquid-phase kinetics based on umbrella sampling.[33] We have also shown that accurate VTST/MT reaction rate coefficients can be obtained using MCMM potentials constructed with as little as one or a few electronic structure Hessians,[31] using standard (e.g., MM3)[43] molecular mechanics force fields. In subsequent work,[34] by testing an MCMM potential for quasiclassical trajectories, we found that, even with relatively good molecular mechanics force fields, it is hard to converge a global potential energy surface to better than 1−2 kcal/mol. In that study, we identified the key limitation of achieving high global accuracy in practical calculations. In particular, we found that the key limitation is the inability of the previous formulation to improve the potential energy surface in regions where the accurate result is higher in energy than the lower of the reactant and product

* Corresponding author e-mail: o_t@t1.chem.umn.edu; truhlar@umn.edu.

10.1021/ct900077g CCC: $40.75
Published on Web 05/21/2009

molecular mechanics approximations. Here, we show that this can be overcome by a non-Hermitian formulation of the theory. The goal of the present work is to show that the new non-Hermitian formulation of the MCMM procedure can be used to fit semiglobal potential energy surfaces with much higher accuracy than the original Hermitian MCMM.

2. Non-Hermitian MCMM

2.1. Key Elements of the New Formulation. In MCMM, the potential energy V at a molecular geometry $\mathbf{x}$ is approximated by the lowest eigenvalue of a valence bond configuration interaction Hamiltonian matrix $\mathbf{H}$, defined by the following:

$$\mathbf{H} = \begin{pmatrix} H_{11}(\mathbf{x}) & \beta(\mathbf{x}) \\ \beta(\mathbf{x}) & H_{22}(\mathbf{x}) \end{pmatrix} \tag{1}$$

where H_{11} and H_{22} are analytical representations of valence bond configurations of the reactant and the product (e.g., molecular mechanics potentials), and β is the approximation to the off-diagonal matrix element, H_{12}. In the original MCMM, β is given by Shepard interpolation of modified Taylor series for H_{12},[27] but in non-Hermitian MCMM, we obtain the square β_o^2 of a zeroth approximation to H_{12}^2 by Shepard interpolation of the unmodified H_{12}^2, and then we write β in terms of β_o. The Shepard interpolation step yields

$$\beta_o^2(\mathbf{x}) = \sum_{k=1}^{N} w_k(\mathbf{x}) T_{12}^2(\mathbf{x}, k) \tag{2}$$

where w_k is a Shepard-interpolation weight function, and each quantity $T_{12}^2(\mathbf{x}, k)$ is a second-order Taylor series of H_{12}^2 at a geometry $\mathbf{x}_k$. In non-Hermitian MCMM, we then approximate β by the following:

$$\beta(\mathbf{x}) = \begin{cases} |\beta_o(\mathbf{x})|; & \beta_o^2(\mathbf{x}) \geq 0 \\ iu|\beta_o(\mathbf{x})|; & \beta_o^2(\mathbf{x}) < 0 \end{cases} \tag{3}$$

where

$$u(\mathbf{x}) = \begin{cases} 1; & \beta_o^2(\mathbf{x}) \geq -\Delta^2/4 \\ \Delta/(2|\beta_o|); & \beta_o^2(\mathbf{x}) < -\Delta^2/4 \end{cases} \tag{4}$$

and

$$\Delta = H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}) \tag{5}$$

There are two key points to emphasize in the above formulation. (i) First, by allowing $\mathbf{H}$ to be non-Hermitian, we broaden the range of geometries for which the MCMM fit can be accurate and thus obtain more accurate representations of potential energy surfaces. The meaning of "broadening" is the inclusion, in addition to the geometries where the true potential is lower than either H_{11} and H_{22}, of all those geometries at which the true potential is above H_{11} or H_{22} (this situation is not typical near a saddle point, but it may be the case when one interpolates a global potential energy surface). At all of these points, the MCMM potential and its first and second derivative were previously set equal to the potential and derivatives of H_{11} or H_{22}, whichever is lower, but the new MCMM formalism allows the improvement of the fitted potential energy surface at such geometries due to H_{12}. (ii) Second, by applying the cutoff function u after the interpolation rather than applying a cutoff function at each Shepard point k, as was done previously, we greatly simplify the algebra, which results in shorter computation times.

The condition $\beta_o^2 < 0$ in eq 3 corresponds to the target potential energy surface being greater than one of the diagonal elements, i.e., than one of H_{11} or H_{22}; the second row of eq 3 allows us to improve the molecular mechanics approximation in this case, and, in fact, we can make MCMM agree exactly with the target data if $\beta_o^2 \geq -\Delta^2/4$.

2.2. Details of Algorithm. The second order Taylor series expansions $T_{12}^2(\mathbf{x}, k)$ used in the Shepard interpolation of eq 2, are constructed in the same way as in steps eqs A5−A11 of the Appendix of ref 34, which is the same as in an earlier version[27] of the MCMM procedure. In particular, we define a matrix $\mathbf{H}^{(k)}$ at each Shepard node k as follows:

$$\mathbf{H}^{(k)} = \begin{pmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{12}^{(k)} & H_{22}^{(k)} \end{pmatrix} \tag{6}$$

Expanding both the diagonal and the off-diagonal elements in Taylor series around a geometry $\mathbf{x}_k$ and using a Taylor series reversion[44] of H_{12}^2, one obtains,

$$T_{12}^2(\mathbf{r}, k) = D^{(k)}\left(1 + \mathbf{b}^{(k)T}\Delta\mathbf{r}^{(k)} + \frac{1}{2}\Delta\mathbf{r}^{(k)T}\mathbf{C}^{(k)}\Delta\mathbf{r}^{(k)}\right) \tag{7}$$

where $T_{12}^2(\mathbf{r}, k)$ is the value at a geometry $\mathbf{r}$ of the expansion of H_{12}^2 in a quadratic Taylor series centered at a Shepard node k; D, b, and C are Taylor series coefficients at that Shepard node k defined by eqs 20−22 of ref 27, and $\Delta\mathbf{r}^{(k)}$ is the difference between the value of a coordinate $\mathbf{r}$ at a given geometry and at Shepard node k. Note that this step and the Shepard interpolation given in eq 2 are performed in internal rather than in Cartesian coordinates $\mathbf{x}$; this set of coordinates (which is called set $\mathbf{r}$ to be consistent with the notation of a previous paper)[32] can be redundant or nonredundant.

The lowest eigenvalue of eq 1, which is an MCMM approximation to the Born−Oppenheimer potential energy, is given by

$$V = \frac{1}{2}(H_{11}(\mathbf{q}(\mathbf{x})) + H_{22}(\mathbf{q}(\mathbf{x})) - [(H_{11}(\mathbf{q}(\mathbf{x})) - H_{22}(\mathbf{q}(\mathbf{x})))^2 + 4\beta^2(\mathbf{r}(\mathbf{x}))]^{1/2}) \tag{8}$$

where β^2 is given by the square of eq 3. Note that β^2 is equal to the quantity β_o^2 obtained directly by Shepard interpolation for all positive β_o^2 and for those negative values β_o^2 that are larger than $-\Delta^2/4$. The analytical first and second derivatives of eq 8 are given in Appendix A. These derivatives involve derivatives of H_{11}, H_{22}, and β^2. The first and second derivatives of H_{11} and H_{22} are calculated analytically by a molecular mechanics code in "natural" internal coordinates $\mathbf{q}$ that are used to express molecular mechanics potentials and are then transformed to Cartesian coordinates. The first and second derivatives of β^2 are calculated from eqs 2−4 in internal coordinates $\mathbf{r}$ and are also transformed to Cartesian coordinates. The gradient and Hessian of V given by eqs 10 and 11 of the Appendix are calculated in Cartesian coordinates. The usage of different sets of coordinates in the MCMM procedure is discussed in detail in ref 32.

The weight function w_k of eq 2 is given by,

$$w_k = \frac{\dfrac{Y_k}{d_k(\mathbf{s})^4}}{\sum_{k=1}^{N+2} \dfrac{Y_k}{d_k(\mathbf{s})^4}} \quad (9)$$

where the variable d_k is a generalized distance[27] between the current geometry and the geometry at Shepard node k, expressed in internal coordinates **s** (we give the internal coordinates used in eq 9 a different name because it is usually convenient to use different coordinates here than were used above). Y_k can be approximated in different ways for different applications. For variational transition state theory calculations (as in the application presented below), where one is only interested in the region of a potential energy surface ranging, in terms of the intermolecular distance between the fragments, from the van der Waals complex of the reactants to the van der Waals complex of the products, one can simply take Y_k as unity. When one requires the potential energy surface at larger intermolecular distances, i.e., beyond the van der Waals complexes of reactants and products, then one can use a function like eq 31 of ref 34.

3. Application to Model Reactions OH + H_2 → H_2O + H and H_2O + H → OH + H_2

The improved MCMM scheme can be applied for constructing semiglobal[45] potential energy surfaces that are invariant under permutations of selected identical nuclei, where such invariance is required (for example, in full-dynamics calculations of gas-phase reactions).[34] This is achieved by properly symmetrizing the diagonal and off-diagonal elements of matrix **H**.[32] It can also be used in VTST/MT calculations that only rely on the knowledge of a potential energy surface in the vicinity of a single reaction swath in which case the nuclear permutation symmetry need not be imposed. When we perform VTST/MT calculations without enforcing the nuclear permutation symmetry, the total number of Shepard points is reduced (e.g., from $m!N$ to N for a reactive system with m low-energy symmetrically equivalent reaction channels), and this reduction results in shorter computation times.

The present application is restricted to the nonsymmetrized potential energy surface of the reactions OH + H_2 → H_2O + H and H_2O + H → OH + H_2, which are used to evaluate the performance of the improved interpolation scheme. We compare the interpolated energies and the VTST/MT rate constants calculated using the MCMM potential energy surfaces to target results calculated directly, i.e., without interpolation. Since our goal is to test an MCMM interpolation method, the use of any electronic structure method that yields a smooth potential energy surface is appropriate. As in the previous work,[34] the target results are obtained using the MPWB1K[46] density functional in conjunction with the 6-31+G(d,p)[47] basis set.

The VTST/MT calculations were performed using the POLYRATE[48] code, the MPWB1K/6-31+G(d,p) energies, gradients, and Hessians were obtained using the *Gaussian*[49] code, the MCMM energies, gradients, and Hessians were obtained using a modified version the MC-TINKER2008-2 code[50] (MC-TINKER2008-2 uses TINKER[51] for molecular mechanics calculations). The VTST/MT calculations on the interpolated surfaces were carried out using the MC-TINKERATE[52] program, which is an interface between MC-TINKER-2008-2 and POLYRATE.

Molecular mechanics force fields used in the present application are given in the Supporting Information. All parameters, except bond dissociation energies, are taken from the previous work.[34] In previous work, all bond dissociation energies were set to values exceeding their accurate values to avoid negative V_{12}^2, but now we use values that are close to experimental values, and this improves the quality of the fit even in regions where one does not place Shepard points.

Two interpolated potential energy surfaces (PES) are considered in the present paper. The first, called PES1, is based on 14 Shepard points, and the second, called PES2 is based on 11 Shepard points. In each case, there are Shepard points at the reactant and product van der Waals minima, at which locations we set $T_{12}^2 = 0$; whereas at the other Shepard points, called electronic structure Shepard points, one obtains T_{12}^2 from MPWB1K calculations. There are 12 electronic structure Shepard points for PES1 and 9 electronic structure Shepard points for PES2.

The electronic structure Shepard points for PES1 are placed at the saddle point optimized at the target level and at 11 nonstationary points. Ten nonstationary points are placed on the minimum energy reaction path (MEP) calculated at the target level, at the following locations: 0.4, 2.0, 3.5, and 4.5 kcal/mol below the saddle point on the OH + H_2 side and 0.2, 0.9, 4.9, 10.1, 14.5, and 17.1 kcal/mol below the saddle point on the H_2O + H side; and one nonstationary point is placed on the concave side of the MEP at a point where the energy is 42 kcal/mol above the H_2O + H asymptote.

The electronic structure Shepard points for PES2 are placed at the saddle point and at 8 nonstationary points. The first nonstationary point is placed on the target level MEP at the point where the energy is 2.0 kcal/mol below the saddle point on the reactant side, and the remaining 7 points were added iteratively, each on the MEP of an MCMM surface with one less electronic structure Shepard points; this is similar to the procedure described in ref 28. These points are located 0.5, 1.1, 2.5, and 2.9 kcal/mol below the saddle point on the OH + H_2 side, and 1.0, 8.0, and 8.6 kcal/mol on the H_2O + H side. The full sets of Cartesian coordinates for all Shepard points of both surfaces are given in Supporting Information.

As explained in Section 2, we use three different internal coordinate sets: set **q** for molecular mechanics calculations, set **r** for Shepard interpolation, and set **s** to calculate the Shepard weighting function. In the present application, the set **r** consists of six nonredundant internal coordinates (three bond distances, two bond angles, and a torsion), and the internal coordinates used to calculate weight function (set **s**) consists of three interatomic distances; all of these coordinates are shown in Figure 1. Previously,[32,34] we only considered cases when set **s** is equivalent to set **r**, but in the

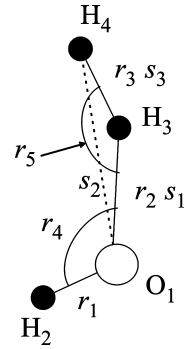


Figure 1. Internal coordinates used in Shepard interpolation (set **r**) and in eq 9 (set **s**).

present work, we introduce some flexibility in the coordinate choice by allowing **s** be different from **r**. Although it was found[27] that an optimal choice for set **s** is often not the same as for set **r** (for example, for VTST/MT calculations of an atom transfer reaction $AX\cdots B \rightarrow A\cdots XB$, a recommended choice for coordinates **s** includes three interatomic distances that undergo significant changes along the MEP, such as, $A - X$, $X - B$, and $A - B$, whereas set **r** requires at least $3N - 6$ nonredundant coordinates), the original implementation[27] of the MCMM procedure lacked the ability to correctly handle this situation. Appendix B gives the Jacobians and Hessians required for MCMM first and second derivatives for an atom transfer reaction with the local coordinates shown in Figure 1. Set **r** is also the same as the set of internal coordinates used to calculate generalized normal mode vibrational frequencies[30,37] along the minimum energy reaction path in VTST/MT calculations.

First, we consider the accuracy of the interpolated energies. In the general case $u(\mathbf{x})$ can be made smooth by joining the two regions of eq 4 by a spline function. In the present, case we simply set $u(\mathbf{x}) = 1$ because none of the geometries involved in the present tests has $\beta_o^2 < -\Delta^2/4$.

We test the surfaces on a 20×20 grid of molecular geometries. The grid was generated by varying the two key bond distances of the transferring H atom while fixing the remaining geometrical parameters (one bond distance, two angles, and a torsion) at their values at the reaction saddle point; the key distances span the ranges $r_2 = 0.78-1.92$ Å and $r_3 = 0.58-1.72$ Å, and total number of geometries is 400. Then, we deleted all energies above 64 kcal/mol with respect to $H_2O + H$; this leaves 338 points. Although these geometries do not span the whole range of dynamically important nuclear configurations, they comprise a representative set of such configurations that may be used to evaluate the accuracy of the fit. Figures 2 and 3 show two-dimensional slices of the target potential energy surface and one of the interpolated ones (PES1) as functions of these two bond distances. Table 1 lists mean unsigned errors for PES1 and PES2 calculated using an old MCMM algorithm[27,34] and using the improved MCMM procedure presented above. The errors are shown as functions of potential energy for four different energy ranges below 64 kcal/mol; the smallest subset of geometries (in the lowest energy range) comprises 46 geometries, and the largest subset comprises 338 geometries. The results indicate that the new MCMM procedure leads to more accurate interpolated potential energies than the Hermitian one, especially in the regions where one places Shepard points. The more accurate fit for PES1 as compared to PES2 is due to the placement of Shepard points in a wider energy range in the former case. In particular, the energy

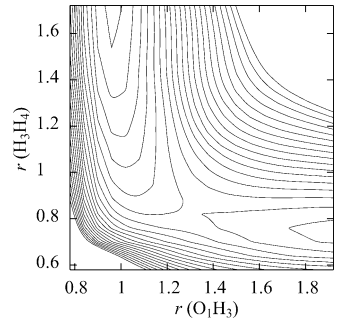


Figure 2. Equipotential contour plot of PES1 as a function of the OH and HH distances. The remaining internal coordinates (one bond distance, two bond angles, and a torsion) are fixed at their values at the reaction saddle point. Contours start at −10.0 kcal/mol and are equally spaced by 3 kcal/mol, with the zero of energy at the OH + H_2 asymptote. Bond distances are in Å.

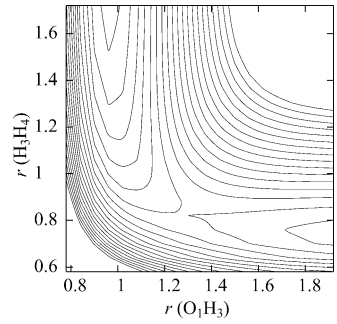


Figure 3. Equipotential contour plot of the target (uninterpolated) potential energy surface. The remaining internal coordinates (one bond distance, two bond angles, and a torsion) are fixed at their values at the reaction saddle point. Contours start at −10.0 kcal/mol and are equally spaced by 3 kcal/mol, with the zero of energy at the OH + H_2 asymptote. Bond distances are in Å.

Table 1. Mean Unsigned Errors[a] in energies for Two Potential Energy surfaces for Various Ranges of Potential energy[b]

energy range	N[c]	PES1 (current algorithm)	PES2 (current algorithm)	PES1 (old algorithm)	PES2 (old algorithm)
<14	46	0.1	0.5	0.6	1.0
<20	116	0.3	0.7	0.6	0.8
<26	177	0.4	0.9	0.7	0.9
<64	338	0.9	1.9	1.2	1.6

[a] In kca/mol. [b] Zero of energy for this table corresponds to H_2O + H. The forward and reverse barrier heights for this reaction with full MPWB1K/6-31+G(d,p) calculations are 4.94 and 18.57 kcal/mol, respectively. Therefore, the number in the first column can be converted to a scale relative to reactants by subtracting 18.57−4.94 = 13.63 kcal/mol. [c] Number of points in this range.

Table 2. Mean Unsigned Percentage Errors in Rate Coefficients for Two Potential Energy Surfaces

T	k^{CVT}	$k^{CVT/ZCT}$	$k^{CVT/SCT}$	$k^{CVT/LCT}$	$k^{CVT/\mu OMT}$
			PES1		
			$OH + H_2 \rightarrow HOH + H$		
300	8	13	3	15	3
400	7	10	5	11	5
600	6	9	7	9	7
			$HOH + H \rightarrow OH + H_2$		
300	7	13	2	14	2
400	6	9	4	10	4
600	4	8	6	8	6
			PES2		
			$OH + H_2 \rightarrow HOH + H$		
300	0	1	10	2	10
400	1	4	2	4	2
600	2	4	1	4	1
			$HOH + H \rightarrow OH + H_2$		
300	1	1	11	1	11
400	0	3	3	3	3
600	1	3	0	3	0

Table 3. Mean Unsigned Percentage Errors Averaged over Three Temperatures (300 K, 400 K, and 600 K) for Reactions $OH + H_2 \rightarrow HOH + H$ and $HOH + H \rightarrow OH + H_2$

T	k^{CVT}	$k^{CVT/ZCT}$	$k^{CVT/SCT}$	$k^{CVT/LCT}$	$k^{CVT/\mu OMT}$
PES1	6	10	5	11	5
PES2	1	3	5	3	5

ranges where the Shepard points are present are 1.5−42.0 kcal/mol for PES1 and 10.0−18.6 kcal/mol for PES2.

Tables 2 and 3 present mean unsigned percentage errors in rate coefficients calculated using the two new MCMM potential energy surfaces. The mean unsigned percentage error is defined as percentage accuracy of the rate coefficient calculated using an MCMM potential as compared to the target result[28,31] obtained via direct dynamics calculations. While the canonical variational transition state rate constant, k^{CVT}, provides a test of the accuracy of the fit near the top of the free energy of activation barrier, the rate coefficients including the tunneling correction (ZCT,[53] SCT,[40,53] LCT,[8] and μOMT)[8,42] depend on the potential in a much wider region. The results for k^{CVT} and for other rate constants are on the average better for PES2 than are the results for PES1. The smaller errors for PES2 are mainly due to the nearly precise location of the variational transition state on this surface because of the availability of a Shepard point in a close proximity to the reaction bottleneck (note that unlike PES1, the PES2 is constructed by placing Shepard points according to a scheme similar to the one in ref 28 particularly designed for VTST/MT calculations, whereas the locations of the data points on PES1 are more or less arbitrary). Excellent results on both interpolated surfaces are obtained for k^{LCT} calculations, which are very sensitive to the potential off the MEP, even though in the case of PES2 no Shepard points are placed on the concave side of the minimum energy reaction path and in the case of PES1 only one point is placed there.

The MCMM scheme is a method for constructing full-dimensional potential energy surfaces for dynamics applications (such as VTST, classical trajectories, or other methods). Although the full advantages of the improved MCMM technique are most apparent in full dynamics simulations, the results of VTST/MT reaction rate calculations also benefit from the improvements introduced the present work in the following ways: (i) the lowest diagonal matrix element may appear to be above the true potential energy surface even at geometries on or near MEP, therefore, using a non-Hermitian matrix **H** improves the accuracy of the MEP on the fitted surface, (ii) the VTST/MT calculations, especially those including large-curvature tunneling, depend on the surface in a much wider range of geometries, therefore, by allowing **H** to be non-Hermitian one improves the overall results, and (iii) eliminating artificial modification in Shepard interpolation introduced in the original MCMM[27] also eliminates sudden changes in the coupling term (caused by forcing $T^2_{\uparrow 2}$ to zero at geometries where is negative) and thus results in a smoother MEP and smoother vibrationally adiabatic ground-state potential curve (obtained by adding zero-point energies of orthogonal harmonic vibrations to the MEP), as well as smoother changes in vibrational frequencies along the reaction path.

Note that H_{11} and H_{22} are expressed here in terms of standard molecular mechanics functions, but one could also replace them with very accurate fits to ab initio data if higher precision in the calculations is sought. Even when H_{11} and H_{22} are represented by standard molecular mechanics, the results in Table 3 indicate that the rate coefficients are converged within 5% of the target results (when one uses the strategy for placing Shepard points designed for VTST calculations,)[28] or within 15% (when the points are placed more or less arbitrarily); in each case these results are within the typical uncertainty of 25% in best estimates of rate coefficients. This implies that MCMM can be used in an automated way (without fitting or adjustments and using a prescription[28] for placing Shepard points) to construct reasonably accurate semi-global representations of potential energy surfaces. However, due to the possibility to replace the standard molecular mechanics picture with more accurate representations for H_{11} and H_{22}, there is enough flexibility to get an arbitrarily high accuracy.

Although the new algorithm is tested here for a tetratomic system, because that allows for well-defined high-precision tests of its ability to fit the dependence on reactive coordinates, we emphasize that MCMM is designed not to compete with modern algorithms for fitting few-body surfaces to spectroscopic accuracy (although, when used with care, it may be competitive with such algorithms) but rather is designed to provide a practical method to fit potential energy surfaces for complex systems.

4. Summary

In this work, we present a new formulation of the multi-configuration molecular mechanics algorithm that improves the accuracy of interpolated potential energy surfaces.

Acknowledgment. This work was supported in part by the NSF under Grant No. CHE07-04974 (dynamics of complex systems, global potential energy surfaces), by the DOE (gas-phase variational transition state theory) under Grant No. CHE07-04974, and by the office of Naval Research (integrated software tools for dynamics) under award No. N00014-05-1-0538.

Supporting Information Available: Force field parameters and geometries used to construct potential energy surfaces and reaction rate coefficients for reactions OH + $H_2 \rightarrow H_2O$ + H and H_2O + H → OH + H_2. This material is available free of charge via the Internet at http://pubs.acs.org.

Appendix A: The Gradient and Hessian of *V*(x) in Cartesian Coordinates

The components of the gradient of an MCMM approximation of the potential energy are given by,

$$G_i = \frac{\partial V}{\partial x_i} = \frac{1}{2}\left(G_{11i}(\mathbf{x}) + G_{22i}(\mathbf{x}) - \left(\frac{2\left(\frac{\partial \beta^2(\mathbf{r}(\mathbf{x}))}{\partial x_i}\right) + (H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))(G_{11i}(\mathbf{x}) - G_{22i}(\mathbf{x}))}{((H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))^2 + 4\beta^2(\mathbf{r}(\mathbf{x})))^{1/2}}\right)\right) \tag{10}$$

where H_{11}, H_{22} are diagonal elements of matrix **H**, G_{11i}, and G_{22i} are components of their gradients, and β is given by eq 3. Note that the first and second derivatives of H_{11} and H_{22} are calculated in internal coordinates **q**(**x**) and then transformed to Cartesian coordinates. The elements of the Hessian of an MCMM approximation of the potential energy are given by,

$$F_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j} = \frac{1}{2}(F_{11ij}(\mathbf{x}) + F_{22ij}(\mathbf{x})) + \frac{\left(2\left(\frac{\partial \beta^2(\mathbf{r}(\mathbf{x}))}{\partial x_i}\right) + (H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))(G_{11i}(\mathbf{x}) - G_{22i}(\mathbf{x}))\right)}{((H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))^2 + 4\beta^2(\mathbf{r}(\mathbf{x})))^{3/2}} \times \frac{\left(2\left(\frac{\partial \beta^2(\mathbf{r}(\mathbf{x}))}{\partial x_j}\right) + (H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))(G_{11j}(\mathbf{x}) - G_{22j}(\mathbf{x}))\right)}{((H_{11}(\mathbf{x}) - H_{22}(\mathbf{x}))^2 + 4\beta^2(\mathbf{r}(\mathbf{x})))^{3/2}} - \frac{(G_{11i} - G_{22i})(G_{11j} - G_{2j})}{((H_{11} - H_{22})^2 + 4\beta^2(\mathbf{r}(\mathbf{x})))^{1/2}} - \frac{2\left(\frac{\partial^2 \beta^2(\mathbf{r}(\mathbf{x}))}{\partial x_i \partial x_j}\right) + (H_{11} - H_{22})(F_{11ij} - F_{22ij})}{((H_{11} - H_{22})^2 + 4\beta^2(\mathbf{r}(\mathbf{x})))^{1/2}} \tag{11}$$

where F_{11ij} and F_{22ij} are the elements of Hessians of H_{11} and H_{22}. The first and second derivatives of β^2 with respect to the coordinates **r** are the same as the derivatives of β_o for all $\beta_o^2 \geq -\Delta^2/4$. These derivatives are given by

$$\mathbf{g} \equiv \frac{\partial \beta_o^2}{\partial \mathbf{r}} = \sum_{k=1}^{N}\left[\frac{\partial w_k}{\partial \mathbf{r}} T_{12}^2(\mathbf{r};k) + w_k D^{(k)}(\mathbf{b}^{(k)} + \mathbf{C}^{(k)}\Delta \mathbf{r}^{(k)})\right] \tag{12}$$

$$\mathbf{f} \equiv \frac{\partial^2 \beta_o^2}{\partial \mathbf{r}^2} = \sum_{k=1}^{N}\left(\frac{\partial^2 w_k}{\partial \mathbf{r}^2} T_{12}^2(\mathbf{r},k) + \frac{\partial w_k}{\partial \mathbf{r}}\mathbf{g}(\mathbf{r}) + D^{(k)}(\mathbf{b}^{(k)} + \mathbf{C}^{(k)}\Delta r^{(k)})\left(\frac{\partial w_k}{\partial \mathbf{r}}\right)^T + w_k D^{(k)}\mathbf{C}^{(k)}\right) \tag{13}$$

where

$$\frac{\partial w_k}{\partial r_\alpha} = \sum_{\gamma=1}^{\Gamma} \frac{\partial w_k}{\partial s_\gamma}\frac{\partial s_\gamma}{\partial r_\alpha} \tag{14}$$

$$\frac{\partial^2 w_k}{\partial r_\alpha \partial r_\beta} = \sum_{\gamma=1}^{\Gamma}\sum_{\gamma=1'}^{\Gamma} \frac{\partial s_\gamma}{\partial r_\alpha}\frac{\partial^2 w_k}{\partial s_\gamma \partial s_{\gamma'}}\frac{\partial s_{\gamma'}}{\partial r_\beta} + \sum_{\gamma=1}^{\Gamma}\frac{\partial w_k}{\partial s_\gamma}\frac{\partial^2 s_\gamma}{\partial r_\alpha \partial r_\beta} \tag{15}$$

where **r** and **s** are the sets of the internal coordinates used in Shepard interpolation and in calculations of the weight function.

Appendix B: Internal Coordinates s and r Used in the Present Application and the Jacobians and Hessians Required by eqs 14 and 15

The internal coordinates **s** and **r** used in eqs 7 and 9 are $\mathbf{r} \equiv \{r_1, r_2, r_3, r_4, r_5, r_6\}$ and $\mathbf{s} \equiv \{s_1, s_2, s_3\}$, respectively; these coordinates are shown in Figure 1. The Jacobians required by eqs 14 and 15 for these coordinates are

$$\frac{\partial s_1}{\partial \mathbf{r}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \frac{\partial s_2}{\partial \mathbf{r}} = \begin{bmatrix} 0 \\ \dfrac{r_2 - r_3\cos(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} \\ \dfrac{r_3 - r_2\cos(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} \\ 0 \\ \dfrac{r_2r_3\sin(r_5)}{r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} \\ 0 \end{bmatrix} \quad \frac{\partial s_3}{\partial \mathbf{r}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{16}$$

The corresponding matrices of second derivatives are as follows:

$$\frac{\partial^2 s_1}{\partial \mathbf{r}^2} = \mathbf{0} \tag{17}$$

$$\frac{\partial^2 s_2}{\partial \mathbf{r}^2} = \begin{bmatrix} 0 & & & & & \\ 0 & \dfrac{1}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & & & & \\ & -\dfrac{1}{2}\dfrac{(r_2 - r_3\cos(r_5))^2}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & & & & \\ 0 & \dfrac{-\cos(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & \dfrac{1}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & & & \\ & -\dfrac{(2r_2 - 2r_3\cos(r_5))(2r_3 - 2r_2\cos(r_5))}{4(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & -\dfrac{1}{2}\dfrac{(r_3 - r_2\cos(r_5))^2}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & & & \\ 0 & 0 & 0 & 0 & & \\ 0 & \dfrac{r_3\sin(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & \dfrac{r_2\sin(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & 0 & \dfrac{r_2r_3\cos(r_5)}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{1/2}} & \\ & -\dfrac{r_2r_3\cos(r_5)(r_2 - r_3\sin(r_5))}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & -\dfrac{r_2r_3\cos(r_5)(r_3 - r_2\sin(r_5))}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & & -\dfrac{(r_2 - r_3\sin(r_5))}{(r_2^2 + r_3^2 + 2r_2r_3\cos(r_5))^{3/2}} & \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{18}$$

$$\frac{\partial^2 s_3}{\partial \mathbf{r}^2} = \mathbf{0} \tag{19}$$

where **0** is a null matrix. Note that eq (18) shows only the lower triangular part of the symmetric Hessian matrix.

References

(1) Ishtwan, J.; Collins, M. A. *J. Chem. Phys.* **1994**, *100*, 8080.

(2) Collins, M. A. *Theor. Chem. Acc.* **2002**, *108*, 313.

(3) Jin, Z.; Braams, B. J.; Bowman, J. M. *J. Phys. Chem. A* **2006**, *220*, 1569.

(4) Xie, D.; Xu, C.; Ho, T.-S.; Rabitz, H.; Lendvay, G.; Lin, S. Y.; Guo, H. *J. Chem. Phys.* **2007**, *126*, 074315.

(5) Dawes, R.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. *J. Chem. Phys.* **2008**, *128*, 084107.

(6) Albu, T. V.; Espinosa-Garcia, J.; Truhlar, D. G. *Chem. Rev.* **2007**, *107*, 5101.

(7) Valiron, P.; Wernli, M.; Favre, A.; Wiesenfeld, L.; Rist, C.; Kedzuch, S.; Noga, J. *J. Chem. Phys.* **2008**, *129*, 134306.

(8) Liu, L.-P.; Lu, D.-h.; González-Lafont, A.; Truhlar, D. G.; Garrett, B. C. *J. Am. Chem. Soc.* **1993**, *115*, 7806.

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

(10) Baldridge, K.; Gordon, M. S.; Steckler, R.; Truhlar, D. G. *J. Phys. Chem.* **1989**, *93*, 5107.

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

(12) Truhlar, D. G.; Duff, J. W.; Blais, N. C.; Tully, J. C.; Garrett, B. C. *J. Chem. Phys.* **1982**, *77*, 764.

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

(14) Espinosa-Garcia, J.; Corchado, J. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1997**, *119*, 9841.

(15) Garrett, B. C.; Koszykowski, M. L.; Melius, C. F.; Page, M. *J. Phys. Chem.* **1990**, *94*, 7096.

(16) Ellingson, B. A.; Pu, J.; Lin, H.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2007**, *111*, 11706.

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

(18) Pu, J.; Truhlar, D. G. *J. Chem. Phys.* **2002**, *116*, 1468.

(19) Corchado, J. L.; Coitiño, E. L.; Chuang, Y.-Y.; Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 2424.

(20) Ruiz-Pernia, J. J.; Silla, E.; Tuñón, I.; Martí, S. *J. Phys. Chem. B* **2006**, *110*, 17663.

(21) *Combined Quantum Mechanical and Molecular Mechanical Methods*; Gao, J., Thompson, M. A., Eds.; ACS Symposium Series 712; American Chemical Society: Washington, DC, 1998.

(22) (a) London, F. *Z. Elektrochem.* **1929**, *35*, 551. (b) Eyring, H.; Polanyi, M. *Naturwissenschaften* **1930**, *18*, 914.

(23) (a) Coulson, C. A.; Danielsson, U. *Ark. Fys.* **1954**, *8*, 239. (b) Janev, R. K.; Radulovic, Z. M. *Phys. Rev. A* **1978**, *17*, 889.

(24) Raff, L. M. *J. Chem. Phys.* **1974**, *60*, 2222.

(25) (a) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218. (b) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523.

(26) Bala, P.; Grochowski, P.; Nowinski, K.; Lesyng, B.; McCammon, J. A. *Biophys. J.* **2000**, *79*, 1253.

(27) Kim, Y.; Corchado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718.

(28) Albu, T. V.; Corchado, J. C.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 8465.

(29) Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 5048.

(30) Lin, H.; Pu, J.; Albu, T. V.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4112.

(31) Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13530.

(32) Tishchenko, O.; Truhlar, D. G. *J. Chem. Theor. Comp.* **2007**, *3*, 938.

(33) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1032.

(34) Tishchenko, O.; Truhlar, D. G. *J. Chem. Phys.* **2009**, *130*, 024105.

(35) Garrett, B. C.; Truhlar, D. G. *J. Chem. Phys.* **1979**, *70*, 1593.

(36) Garrett, B. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1979**, *101*, 4534.

(37) Jackels, C. F.; Gu, Z.; Truhlar, D. G. *J. Chem. Phys.* **1995**, *102*, 3188.

(38) Truhlar, D. G.; Isaacson, A. D.; Skodje, R. T.; Garrett, B. C. *J. Phys. Chem.* **1982**, *86*, 2252. . **1983**, *87*, 4554 (E).

(39) Truhlar, D. G.; Gordon, M. S. *Science* **1990**, *249*, 491.

(40) Liu, Y.-P.; Lynch, G. C.; Truong, T. N.; Truhlar, D. G.; Garrett, B. C. *J. Am. Chem. Soc.* **1993**, *115*, 2408.

(41) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771.

(42) Fernández-Ramos, A.; Truhlar, D. G. *J. Chem. Phys.* **2001**, *114*, 1491.

(43) Allinger, N. C.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1998**, *111*, 8551.

(44) Chang, Y. T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884.

(45) When referring to MCMM potential energy surfaces, the term "global" indicates that such a surface is defined everywhere in the nuclear configuration space, whereas "semiglobal" is used to emphasize that it is aimed to represent an accurate fit in dynamically relevant regions. For example, in the application described in the present work, the reaction asymptote with all four atoms of the OH + H_2 system separated apart is not neccesserily well represented in the present fit, but because of the high energy of such configuration, this limitation does not affect dynamics in the energy range of interest.

(46) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(47) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(48) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Villa, J.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Melissas, V. S.; Lynch, B. J.; Rossi, I.; Coitiño, E. L.; Fernández-Ramos, A.; Pu, J.; Albu, T. V.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *Polyrate 9.7*; University of Minnesota: Minneapolis, MN, 2007.

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

(50) Tishchenko, O.; Higashi, M.; Albu, T. V.; Corchado, J. C.; Kim, Y.; Villà, J.; Xing, J.; Lin, H. Truhlar, D. G. *MC-TINKER-2008-2*; University of Minnesota: Minneapolis, MN, 2008.

(51) Ponder, J. W. *TINKER-Version 4.2*; Washington University: St. Louis, MO, 2004.

(52) Albu, T. V.; Corchado, J. C.; Kim, Y.; Villa, J.; Xing, J.; Lin, H.; Tishchenko, O.; Higashi, M.; Truhlar, D. G. *MC-TINKERATE-2008*; University of Minnesota: Minneapolis, MN, 2008.

(53) Lu, D.-H.; Truong, T. N.; Melissas, V. S.; Lynch, G. C.; Liu, Y.-P.; Garrett, B. C.; Steckler, R.; Isaacson, A. D.; Rai, S. N.; Hancock, G. C.; Lauderdale, J. G.; Joseph, T.; Truhlar, D. G. *Comput. Phys. Commun.* **1992**, *71*, 235.

CT900077G

JCTC Journal of Chemical Theory and Computation

Thermodynamic Properties of Liquid Water: An Application of a Nonparametric Approach to Computing the Entropy of a Neat Fluid

Lingle Wang, Robert Abel, Richard A. Friesner, and B. J. Berne*

Department of Chemistry, Columbia University, New York, New York 10027

Received February 13, 2009

Abstract: Because of its fundamental importance to molecular biology, great interest has continued to persist in developing novel techniques to efficiently characterize the thermodynamic and structural features of liquid water. A particularly fruitful approach, first applied to liquid water by Lazaridis and Karplus, is to use molecular dynamics or Monte Carlo simulations to collect the required statistics to integrate the inhomogeneous solvation theory equations for the solvation enthalpy and entropy. We here suggest several technical improvements to this approach, which may facilitate faster convergence and greater accuracy. In particular, we devise a nonparametric *k*th nearest-neighbors (NN)-based approach to estimate the water−water correlation entropy, and we suggest an alternative factorization of the water−water correlation function that appears to more robustly describe the correlation entropy of the neat fluid. It appears that the NN method offers several advantages over the more common histogram-based approaches, including much faster convergence for a given amount of simulation data; an intuitive error bound that may be readily formulated without resorting to block averaging or bootstrapping; and the absence of empirically tuned parameters, which may bias the results in an uncontrolled fashion.

1. Introduction

Water is unique among liquids for its biological significance. It plays an active role in the formation of the structures of proteins, lipid bilayers, and nucleic acids in vivo, both through direct hydrogen-bonding interactions with these biomolecules, and also through indirect interactions, where the unique hydrogen-bonded structure of liquid water is known to drive hydrophobic assembly.[1] It has been suggested that a robust characterization of the thermodynamic properties and structure of water solvating the active site of a protein is essential to rationalize the various binding affinities of small molecules that will displace that solvent to bind to the protein active site.[2,3]

As such, great interest has continued to persist in developing novel techniques to efficiently characterize the thermodynamic and structural features of liquid water in different environments. A particularly fruitful approach, first applied to liquid water by Lazaridis and Karplus,[4−6] used molecular dynamics or Monte Carlo simulations to collect the required statistics to integrate the inhomogeneous solvation theory (IST) equations for the solvation enthalpy and entropy. In this theory, the solvation enthalpy is determined from an analysis of the change in the solute−solvent and solvent−solvent interaction energy terms, and the solvation entropy is computed from an expansion of the entropy in terms of increasingly higher order solute−solvent correlation functions.[4] This approach has been used to characterize the thermodynamics and structure of neat water,[6] hydration of small hydrophobes,[4] and the hydration of the active sites of proteins.[7,8] Recently, it has also been extended to allow for the rapid computation of the relative binding affinities of a set of congeneric ligands with a given protein, via a semiempirical displaced-solvent functional.[2]

Because of the increasing interest in applying this technique to water[9−12] in various environments, we have chosen to reexamine the factorization and correlation function integration scheme originally suggested by Lazaridis and Karplus[6] for bulk water and later adopted by

* Corresponding author e-mail: bb8@columbia.edu.

10.1021/ct900078k CCC: $40.75
Published on Web 05/18/2009

others.[13] We have found that several technical improvements in this scheme are possible, which may facilitate faster convergence and greater accuracy than the more typical expressions. In this Article, we (1) devise a nonparametric *k*th nearest-neighbors (NN)[14]-based approach to estimate the water−water correlation entropy, in lieu of the more common histogram-based approaches, and (2) suggest an alternative factorization for the water−water correlation function that appears to more robustly describe the water−water correlation entropy of the neat fluid. To our knowledge, this is the first application of the NN method to compute the entropy of a neat fluid. It appears that the NN method offers several advantages over the more common histogram-based approaches, including (1) much faster convergence for a given amount of simulation data, especially when the correlation function is highly structured; (2) an intuitive error bound may be readily formulated without resorting to block averaging or bootstrapping techniques, which may be problematic to apply to estimators of the entropy; and (3) the absence of empirically tuned parameters, such as the histogram bin width, which may bias the results in an unpredictable fashion. Our alternative factorization of the water−water correlation function explicitly includes correlations between the water−dipole-vector−intermolecular-axis angle with the angle of rotation of the water molecule about its dipole vector. This contribution, although neglected by others,[6] has been found in our work to increase the agreement of results obtained by the entropy expansion with those obtained by less approximate methods, such as free energy perturbation theory. We also extensively compare the solvation entropies obtained from the truncated entropy expansion to those obtained from a finite difference analysis of free energy perturbation theory results. This comparison allows us to characterize the errors in both precision and accuracy associated with the NN method of integrating the entropy expansion presented here.

Our primary interest in developing this technique was to later adapt the method to study the solvation of solutes; thus, we were interested in determining realistic estimates of the convergence of the technique when the isotropic symmetry of the fluid was not present. As such, when extracting the solvent configurations to compute the pair correlation function (PCF), we chose to use only the configurations of a distinguished solvent molecule with the rest of the system, instead of collecting statistics from all pairs of solvent molecules. Such a protocol allows for an interrogation of the relative convergence properties of the various methods that might be obscured by the additional statistics offered by taking advantage of the symmetry of the system.

2. Methods

2.1. The Entropy Expression of a Neat Fluid. First derived by Green,[15] and later by Raveché[16] and Wallace,[17] the entropy of a fluid can be expressed as a sum of integrals over multiparticle correlation functions. For a molecular fluid,[5] the expression is

$$\begin{aligned} s = s^{\mathrm{id}} + s_{\mathrm{e}} = s^{\mathrm{id}} - \frac{1}{2!}k\frac{\rho}{\Omega^2} \\ \int [g^{(2)} \ln(g^{(2)}) - g^{(2)} + 1]\, \mathrm{d}\boldsymbol{r}\, \mathrm{d}\omega^2 \\ -\frac{1}{3!}k\frac{\rho}{\Omega^2}\int [g^{(3)} \ln(\delta g^{(3)}) - g^{(3)} + \\ 3g^{(2)}g^{(2)} - 3g^{(2)} + 1]\, \mathrm{d}\boldsymbol{r}_1\, \mathrm{d}\boldsymbol{r}_2\, \mathrm{d}\omega^3 - \ldots \end{aligned} \tag{1}$$

where, s^{id} is the entropy of an ideal gas with the same density and temperature as the fluid, s_{e} is the excess entropy of the fluid over that of the ideal gas, k is Boltzmann's constant, ρ is the number density, ω denotes the orientational variables of one molecule, Ω is the total volume of the orientational space (for a nonlinear molecule like water, Ω is $8\pi^2$), $g^{(2)}$ is the pair correlation function, $g^{(3)}$ is the triplet correlation function, and $g^{(3)}$ is the deviation of $g^{(3)}$ from the superposition approximation. In practice, it is very difficult or even impossible to converge the three-particle and higher order correlation terms. However, it has been established that, for most fluids, the largest contribution to the excess entropy comes from the two-particle correlation term,[6] and the error induced by neglecting the higher order terms of the expansion may often be safely ignored.

Following the work of Lazaridis and Karplus,[6] we evaluate the two-particle excess entropy of liquid water by separating the two-particle term into translational and orientational components by factorization:

$$g(r, \omega^2) = g(r)g(\omega^2|r) \tag{2}$$

$$s_{\mathrm{e}}^{(2)} = s_{\mathrm{trans}}^{(2)} + s_{\mathrm{orient}}^{(2)} \tag{3}$$

$$s_{\mathrm{trans}}^{(2)} = -\frac{1}{2}k\rho\int [g(r)\ln g(r) - g(r) + 1]\, \mathrm{d}\boldsymbol{r} \tag{4}$$

$$s_{\mathrm{orient}}^{(2)} = \frac{1}{2}k\rho\int g(r)S^{\mathrm{orient}}(r)\, \mathrm{d}\boldsymbol{r} \tag{5}$$

$$S^{\mathrm{orient}} = -\frac{1}{\Omega^2}\int J(\omega^2)g(\omega^2|r)\ln g(\omega^2|r)\, \mathrm{d}\omega^2 \tag{6}$$

where r is the oxygen−oxygen distance of two water molecules, ω^2 are the angles that define the relative orientation of the two water molecules, $J(\omega^2)$ is the Jacobian of the angular variables, $g(r,\omega^2)$ is the pair correlation function, and $g(\omega^2|r)$ is the conditional-angular pair correlation function in the typical Bayesian notion. (Note that $g(r,\omega^2)$ is identical to $g^{(2)}$ as it appears in eq 1.) We denote the relative orientation of the two water molecules by the five angles[6] $[\theta_1,\theta_2,\phi,\chi_1,\chi_2]$, where θ_1,θ_2 are the angles between the intermolecular axis and the dipole vector of each molecule, ϕ describes the relative dihedral rotation of the dipole vector around the intermolecular axis, and χ_1,χ_2 describe the rotation of each molecule around its dipole vector. In the following discussion, we denote the entropy defined by formula 6 the orientational Shannon entropy,[18] and denote the entropy defined by formula 5 the orientational excess entropy.

In line with prior work,[6] we calculated the orientational Shannon entropy as defined by formula 6 for three different ranges of r: $(0 < r \leq 2.7)$, $(2.7 < r \leq 3.3)$, and $(3.3 < r \leq 5.6)$, which correspond to the various peaks and troughs in

the radial distribution function. In this way, the orientational excess entropy is related to Shannon entropy by:

$$s_{\text{orient}} = \frac{1}{2}N_i k S^{\text{orient}} \quad i = 1, 2, 3 \tag{7}$$

where N_i is the average number of water molecules in the ith shell.

2.2. Factorization of the Orientational Pair Correlation Function Using Generalized Kirkwood Superposition Approximation. The orientational pair correlation function (PCF) of water is a function of five angles, which is very difficult to converge from currently accessible molecular dynamics simulation time scales. The idea of factorization is to approximate the higher dimensional probability density function by the product of its lower dimensional marginal probability density functions. The generalized Kirkwood superposition approximation (GKSA)[19−21] allows an m-dimensional distribution to be estimated using corresponding $m - 1$-dimensional distributions:

$$\rho(x_1, x_2, ..., x_m) = \begin{cases} \dfrac{\prod_{c_{m-1}^m} \rho_{m-1} \cdots \prod_{c_2^n} \rho_2}{\prod_{c_{m-2}^m} \rho_{m-2} \cdots \prod_{c_1^n} \rho_1} & m \text{ is odd} \\ \dfrac{\prod_{c_{m-1}^m} \rho_{m-1} \cdots \prod_{c_1^n} \rho_1}{\prod_{c_{m-2}^m} \rho_{m-2} \cdots \prod_{c_2^n} \rho_2} & m \text{ is even} \end{cases} \tag{8}$$

where ρ_{m-k} represents a specific probability density function of $m - k$ dimensionality, and c_{m-k}^m indicates all possible combinations of $m - k$ groupings from the set of m total variables. Reiss[20] and Singer[21] have demonstrated that the GKSA is the optimal approximation of an n-particle distribution for $n \geq 3$ from a variational point of view, and it has been applied in numerous settings.[22,23]

From the results of our simulations, and as indicated by Lazaridis and Karplus,[6] the distribution has no structure along angle ϕ; that is, $g(\phi)$ is close to 1 over the range of ϕ and has no correlation with other angles. Thus, we approximated the five-dimensional PCF by:

$$g(\theta_1, \theta_2, \phi, \chi_1, \chi_2) = g(\theta_1, \theta_2, \chi_1, \chi_2) g(\phi) \tag{9}$$

Note that for any properly defined orientational PCF $g(x_1,x_2,...x_N)$,

$$\frac{1}{\Omega^{[x_1,x_2...x_n]}} \int J(x_1, x_2...x_n) g(x_1, x_2...x_n)\, dx_1\, dx_2...dx_n = 1 \tag{10}$$

where

$$\Omega^{[x_1,x_2...x_n]} = \int J(x_1, x_2...x_n)\, dx_1\, dx_2...dx_n \tag{11}$$

That is, $\Omega^{[x_1,x_2,...x_n]}$ is the integral of the Jacobian $J(x_1,x_2,...x_n)$ over angular variables $x_1,x_2,...x_n$. Therefore, $g(x_1,x_2,...x_n)$ is proportional to $\rho(x_1,x_2,...x_n)$ with proportional coefficient $\Omega^{[x_1,x_2,...x_n]}$. Via application of the GKSA (formula 8), it follows:

$$g(\theta_1, \theta_2, \chi_1, \chi_2) = \frac{g(\theta_1, \theta_2) g(\theta_1, \chi_1) g(\theta_1, \chi_2) g(\theta_2, \chi_1) g(\theta_2, \chi_2) g(\chi_1, \chi_2)}{g^2(\theta_1) g^2(\theta_2) g^2(\chi_1) g^2(\chi_2)} \tag{12}$$

Note that this factorization differs from that introduced by Karplus and Lazaridis[6] by the explicit inclusion of $g(\theta_1,\chi_1)$ and $g(\theta_2,\chi_2)$ terms. Taking this approximation of $g(x_1,x_2,...x_n)$ into the argument of the logarithm of formula 6, we find

$$S^{\text{orient}} = -\frac{1}{\Omega^2} \int J(\omega^2) g(\omega^2|r) \ln g(\omega^2|r)\, d\omega^2 \tag{13}$$

$$= -\sum_{C_2^4} \frac{1}{\Omega^{[x_1,x_2]}} \int J(x_1, x_2) g(x_1, x_2) \ln g(x_1, x_2)\, dx_1\, dx_2 + 2 \sum_{C_1^4} \frac{1}{\Omega^{[x]}} \int J(x) g(x) \ln g(x)\, dx \tag{14}$$

$$= \sum_{C_2^4} S^{[x_1,x_2]} - 2 \sum_{C_1^4} S^{[x]} \tag{15}$$

where x_1,x_2 is any combination of two variables from the $[\theta_1,\theta_2,\chi_1,\chi_2]$ set, x is any variable from the $[\theta_1,\theta_2,\chi_1,\chi_2]$ set, $J(x_1,x_2)$ is the Jacobian of the corresponding two variables, $J(x)$ is the Jacobian corresponding to variable x, $\Omega^{[x_1,x_2]}$ is the total accessible angular volume of variables x_1,x_2, $\Omega^{[x]}$ is the total accessible angular volume of variable x, $S^{[x_1,x_2]}$ is the Shannon entropy of angular variables x_1 and x_2, and $S^{[x]}$ is the Shannon entropy of angular variable x.

We note that an ambiguity seems to exist in the literature as to how to properly apply an approximation of the type suggested in eq 12 to eq 6. We have adopted here to apply the approximation only to the logarithm of eq 6 (as was done in the original derivation of eq 1), which allows result 15 to be interpreted through the language of information theory.[24] An alternate approach, which has been adopted by others, has been to apply approximation 12 to both occurrences of the PCF in eq 6, taking care to renormalize the factorization of the PCF introduced in eq 12 so that meaningful results will still be obtained. Interestingly, the results of these two approaches do not numerically agree, which may not be obvious from cursory inspection. We leave this proof as an exercise for the reader, which can be readily shown for instance from a correlated multidimensional Gaussian distribution.

2.3. The *k*'th Nearest-Neighbor Method. The NN method[14] gives an asymptotically unbiased estimate of an integral of the form:

$$I = -\int \rho(x_1, x_2, ..., x_s) \ln \rho(x_1, x_2, ..., x_s)\, dx_1\, dx_2...dx_s \tag{16}$$

where $\rho(x_1,x_2,...x_s)$ is the probability density function. Given a reasonable estimation of probability density function $f(x^i)$, the value of integral can be approximated as

$$I \approx -\frac{1}{n} \sum_{i=1}^{n} \ln f(x^i) \tag{17}$$

which follows from x^i being sampled from the true distribution $\rho(x^i)$. The NN method of nonparametrically estimating $f(x^i)$ at a point $x^i = (x_1^i,x_2^i,...x_s^i)$ is[25]

$$f(x^i) = \frac{k}{n}\frac{1}{V_s(R_{i,k})} \tag{18}$$

$$V_s(R_{i,k}) = \frac{\pi^{s/2}R_{i,k}^s}{\Gamma\left(\frac{1}{2}s + 1\right)} \tag{19}$$

where n is the number of data points in the sample, $V_s(R_{i,k})$ is the volume of an s-dimensional sphere with radius $R_{i,k}$, and $R_{i,k}$ is the Euclidean distance between the point x^i and its kth nearest neighbor in the sample. This approximation amounts to assuming that the distance between neighboring sampled points in configuration space will be small where the probability density function is large, and vice versa. So this integration may be estimated as

$$I \approx -\frac{1}{n}\sum_{i=1}^{n}\ln f(x^i) = \frac{1}{n}\sum_{i=1}^{n}\ln\frac{n\pi^{s/2}R_{i,k}^s}{k\Gamma\left(\frac{1}{2}s + 1\right)} \tag{20}$$

However, the estimate in eq 20 is systematically biased[14] and will deviate from the correct result in the limit of large n by $L_{k-1} - \ln k - \gamma$, where $L_j = \Sigma_{i=1}^{j}1/i$ and $\gamma = 0.5772...$ is Euler's constant. By subtracting the bias $L_{k-1} - \ln k - \gamma$, the modified unbiased estimate is formulated as

$$I \approx \frac{s}{n}\sum_{i=1}^{n}\ln R_{i,k} + \ln\frac{n\pi^{s/2}}{\Gamma\left(\frac{1}{2}s + 1\right)} - L_{k-1} + \gamma \tag{21}$$

Now our goal is to modify our expressions for the Shannon entropies into a form that is amenable to a kth NN evaluation of the integral. The expression of the two-dimensional orientational Shannon entropy has the form of

$$S^{[x_1,x_2]} = -\frac{1}{\Omega^{[x_1,x_2]}}\int J(x_1,x_2)g(x_1,x_2)\ln g(x_1,x_2)\,dx_1\,dx_2 \tag{22}$$

where $J(x_1,x_2)$ is the Jacobian associated with x_1 and x_2. Here, for χ_1 and χ_2 the Jacobian is 1, but for θ_1 and θ_2 the Jacobian is sin θ_1 and sin θ_2. However, by a change of variables from θ to $t = \pi/2(\cos\theta + 1)$, the Jacobian for t becomes 1, and the total angular volume is π for one-dimensional distribution and π^2 for two-dimensional distributions. Next, $g(x_1,x_2)$ is proportional to $\rho(x_1,x_2)$ in eq 16, with proportional coefficient π^2. Following the NN method, the statistically unbiased estimation of the one- and two-dimensional orientational Shannon entropies may now be approximated as

$$H_k^{[x]}(n) = \frac{1}{n}\sum_{i=1}^{n}\ln R_{i,k} + \ln\frac{n\pi^{1/2}}{\Gamma\left(\frac{1}{2} + 1\right)\Omega^{[x]}} - L_{k-1} + \gamma \tag{23}$$

$$H_k^{[x_1,x_2]}(n) = \frac{2}{n}\sum_{i=1}^{n}\ln R_{i,k} + \ln\frac{n\pi^{1}}{\Gamma\left(\frac{1}{2}\times 2 + 1\right)\Omega^{[x_1,x_2]}} - L_{k-1} + \gamma \tag{24}$$

where $H_k^{[x]}(n)$ is the kth NN estimate of the Shannon entropy of random variable x from a sampling of n data points, and $H_k^{[x_1,x_2]}(n)$ is the kth NN estimate of the joint Shannon entropy of random variables x_1,x_2 from a sampling of n data points. Thus, we are now equipped to apply the NN method of estimating the entropy to liquid state problems. We also note that to compute the NN distances, we made use of the ANN code,[26] which utilizes the $k-d$ tree algorithm[27] for obtaining the kth NN distances $R_{i,k}$ between sample points as necessary.

2.4. Error Analysis of the *k*th Nearest-Neighbor Method. It has been shown through an analysis of the limiting distribution[14] that the variance of the kth NN estimate of the entropy $H_k(n)$ is

$$\mathrm{var}[H_k(n)] = \frac{Q_k + \mathrm{var}[\ln f(x)]}{n} \tag{25}$$

where $f(x)$ is the probability density function and $Q_k = \sum_{j=k}^{\infty}1/j^2$. Formally, this result follows from using the Poisson approximation of the binomial distribution to characterize the fluctuations of $H_k(n)$ in the large n limit (please see ref 14 for details). Because $H_k(n)$ is asymptotically unbiased,[14] the asymptotic mean square error of the estimate is of the order given by eq 25. Typically, the true value $H(n)$ will be estimated by computing $H_k(n)$ for several values of k, typically 1−5. Because the analytical form of the variance is known, we may combine these estimates by a weighted averaging procedure, that is, $H(n) = \sum w_k H_k(n)$. For independent variables with the same average, the weight that minimizes the variance of the estimate of the average is a weight proportional to the inverse of the variance of the variable (see Appendix A for details), that is,

$$w_k = \frac{1/(Q_k + \mathrm{var}[\ln f(x)])}{\sum_{i=1}^{m}1/(Q_k + \mathrm{var}[\ln f(x)])}\text{for } k = 1, 2...m \tag{26}$$

where w_k is the ideal weight of $H_k(n)$ when averaging $H(n)$. Such calculations may also be readily extended to compute the standard deviation of such an estimate (Appendix A). Interestingly, two well-defined limits exist here: (1) if var[ln $f(x)$] is small, then the proper weighting will be

$$w_k = \frac{1/Q_k}{\sum_{k=1}^{m}1/Q_k}\text{for } k = 1, 2...m \tag{27}$$

and, (2) if var[ln $f(x)$] is large, then the proper weighting will be a flat function, which will lead to a simple arithmetic average. Therefore, the best possible estimate of $H(n)$ from m estimates of $H_k(n)$ will always be bound by these two limiting averages. Further, if these two limiting averages converge in the given sampling, it is highly probable the estimate of $H(n)$ is also converged. We also note here that an intuitive sense of which regime best fits the given data can be discerned by inspecting the relative noise in plots of the m $H_k(n)$ estimates as a function of n (where n is the amount of simulation time in this application). If the $H_1(n)$ estimate noticeably suffers greater fluctuations than the other estimates, then the var[ln $f(x)$] term must be small, because the Q_1 component is dominating relative variances of the estimates. However, if the m $H_k(n)$ estimates all appear

graphically to have fluctuations of a similar magnitude, then the var[ln $f(x)$] term must be large, and the simple arithmetic average is more appropriate. Such inspection of our data revealed var[ln $f(x)$] to be small. As such, the weighted average determined by application of eq 27 was taken in this work as our best possible estimate of $H(n)$.

2.5. Calculation of the Excess Energy, Enthalpy, and Free Energy. The excess molar energy of a fluid is simply

$$\Delta E = \frac{1}{2}\frac{\rho}{\Omega^2}\int g(r,\omega^2)u(r,\omega^2)\,\mathrm{d}r\,\mathrm{d}\omega^2 \qquad (28)$$

where $u(r,\omega^2)$ is the interaction energy between two molecules with distance r and orientation determined by ω^2. This quantity is straightforward to extract from the simulation, as it is merely one-half of the interaction energy between the water molecule of interest with the rest of the system. The molar excess enthalpy can be obtained by approximating the $\Delta(PV)$ term. For the liquid phase, the PV term may be safely neglected, and for the gas phase, we may use the ideal gas equation of state $PV = NkT$ to derive an excellent approximation to the PV term analytically. Combined with the excess entropy, we find the excess free energy of the fluid may be expressed as

$$\Delta G = \Delta E + \Delta(PV) - Ts_e \qquad (29)$$

as is typical.

2.6. The Finite-Difference Method of Entropy Calculation. To generate reference data to examine the accuracy of the kth NN method of evaluating the entropy expansion, we pursued a finite difference analysis of the solvation free energy, as computed from free energy perturbation theory (FEP). The finite-difference (FD) method of computing an entropy from FEP data proceeds by first noting that the entropy is the temperature derivative of the free energy, and then attempting to accurately estimate this slope,[28] that is

$$-\Delta S(T) = \left\langle\frac{\partial\Delta G}{\partial T}\right\rangle_P = \frac{\Delta G(T+\Delta T) - \Delta G(T-\Delta T)}{2\Delta T} \qquad (30)$$

This method relies on the assumption that the heat capacity of the system is independent of temperature in the range $[T - \Delta T, T + \Delta T]$.[29] This assumption appears to be valid near room temperature with ΔT even as large as 50 K.[28] Here, we use the Bennett acceptance ratio[30] method to calculate the excess free energy of liquid water at $T = 298 \pm 20$ K, and then use FD to calculate the excess entropy at $T = 298$ K. The datails of this method are included in the appendices. These data allow for independent validation of the NN approach and the approximations therein.

2.7. Details of the Simulation. Dynamics trajectories were generated using the Desmond molecular dynamics program.[31] A 25 Å cubic box of the TIP4P[32] water model was first equilibrated to 298 K and 1 atm with Nose−Hoover[33,34] temperature and Martyna−Tobias−Klein[35] pressure controls, followed by 30 ns NVT dynamics simulation with a Nose−Hoover[33,34] temperature control. To integrate the equations of motion of the system, the RESPA[36] integrator was used, where the integration step was 2 fs for the bonded and the nonbonded-near interactions and 6 fs for the nonbonded-far interactions. Configurations were collected every 1.002 ps. The cutoff distance was 9 Å for the van der Waals interaction, and the particle-mesh Ewald[37] method was used to model the electrostatic interactions. Similar simulations were performed for the SPC,[38] SPC/E,[39] TIP3P,[32] and TIP4P-Ew[40] water models.

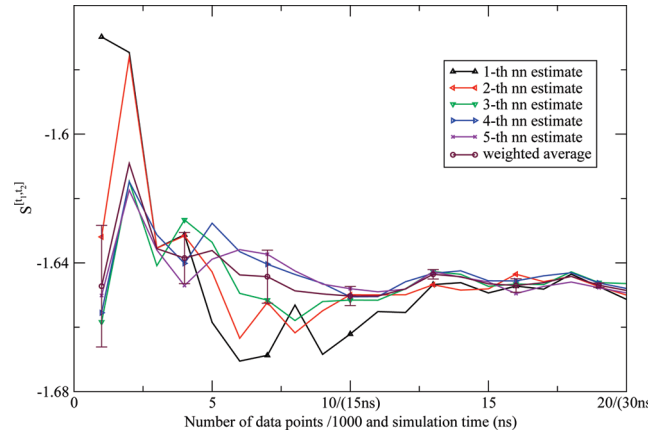


Figure 1. The first shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 1000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using the NN method. The weighted average estimate and the associated error bar were also depicted.

When extracting the solvent configurations to compute the PCF, we chose to only use the configurations of a distinguished solvent molecule with the rest of the system, instead of collecting statistics from all pairs of solvent molecules. Our primary interest in developing this technique was to later adapt the method to study the solvation of solutes; thus, we were interested in determining realistic estimates of the convergence of the technique when the isotropic symmetry of the fluid was not present. Such a protocol allows for an interrogation of the relative convergence properties of the various methods that might be obscured by the additional statistics offered by taking advantage of the symmetry of the system.

3. Results and Discussion

3.1. The Shannon Entropies. The NN estimates of the two-dimensional orientational Shannon entropies $S^{[r_1,r_2]}$ of the TIP3P water model for the three shells are given in Figures 1, 2, and 3. The results reported in these figures were generally representative of those results obtained for the other models. We see from the figures that the weighted average estimate of all of the Shannon entropies is converged over the course of the simulations. The results of all of the one- and two-dimensional orientational Shannon entropies for each of the three shells for all of the water models studied are given in Table 1. By application of formulas 4 and 7, we computed the translational excess entropies and orientational excess entropies for all of the water models studied. All of the final results are shown in Table 2. From the table, we

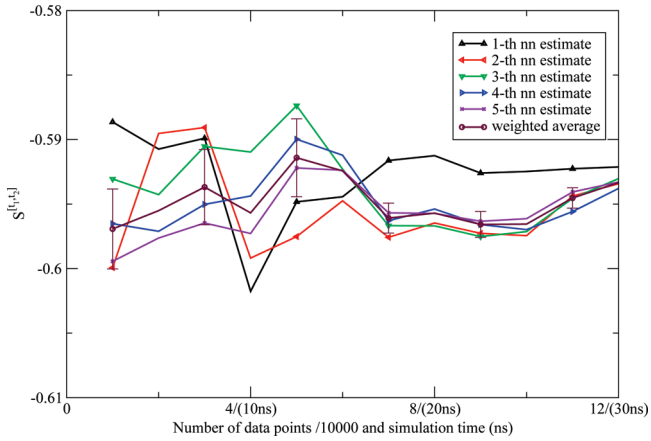


Figure 2. The second shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 10 000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using the NN method. The weighted average estimate and the associated error bar were also depicted.

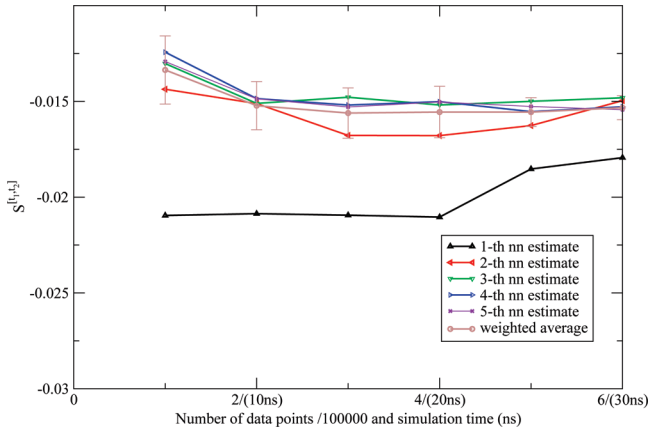


Figure 3. The third shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 100 000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using the NN method. The weighted average estimate and the associated error bar were also depicted.

see that for the TIP4P model the excess entropy result from the NN method, −13.67 eu, is very close to the experimental value, −14.1 eu. We also note excellent agreement between the excess entropies computed here and those derived from cell theory.[41] The agreement for the TIP3P and SPC models was slightly diminished as compared to the other models, for reasons that will be explained later.

3.2. Convergence Properties. We extensively compared the commonly employed histogram method to compute the orientational Shannon entropy to the NN method weighted average (Figures 4, 5, and 6). We see clearly that the NN method weighted average converges much faster than the histogram method for shells 1 and 2. For shell 3, both methods give similar results. This is easily understood: for the first and second shells, the water molecules are highly correlated, and the histogram results will have a strong dependency on the bin size used to do the integration; however, for the third shell, there is little correlation, so the histogram method has similar convergence properties as compared to the NN method.

Figures 7, 8, 9, 10, and 11 depict the total orientational excess entropies as a function of simulation time from the various histogram estimates and the NN weighted average estimate. For all of the models studied, the 10° histogram estimate (which is most commonly used currently[6,10]) gave results closest to the NN estimate. However, for a bin size of 20°, the entropy result is biased away from the correct result, and for bin sizes of 5° and 2.5°, much longer simulation time would be needed to converge the results. Because ideal bin size is problem specific, it cannot be deduced unless other reference data are already known. Thus, the absence of such a parametric bias in the NN method is a notable advantage of the technique.

3.3. Error Analysis. As described in the Methods, we calculated the variance associated with the weighted average of the NN estimates for each of the one- and two-dimensional Shannon entropies. Because the NN estimate is asymptotically unbiased, the error of the estimate is also given by the variance. We calculated the error on the basis of the weighted average, which assumes var ln $f(x)$ is 0. However, even in the extreme cases where var ln $f(x)$ goes to infinity and the five NN estimates contribute equally to the average, the variance of the arithmetic average only differs slightly from weighted average, and they are within the error bar of each other, strongly indicating the convergence of these calculations (Figures 12 and 13).

3.4. The Radial Dependence of Orientational Shannon Entropy. We calculated the orientational Shannon entropies in three radial regions, assuming the orientational distribution would be independent of r in each subregion. To validate this approximation, we calculated the orientational Shannon entropies at different intervals of r from 2.5 to 4.0 Å. Typical Shannon entropies $S^{[t_1,t_2]}$ at different values of r are shown in Figure 14.

We see from the figure that the Shannon entropy increases as the distance between the two water molecules r increases, and goes to zero when r is sufficiently large. Additionally, the change of the Shannon entropy with respect to r is smooth in the respective first and second hydration shells. Because of the slow variation of the orientational Shannon entropy with respect to r, the sum of the orientational excess entropy at each interval will differ from the sum of the orientational excess entropy of the three shells only by at most 0.5 eu, which is within statistical uncertainty of the calculation. Thus, this approximation was not a large source of error in these calculations.

3.5. Inclusion of $g(\theta_1,\chi_1)$ in the Factorization. The factorization of the PCF used here differs from the more common formulation[6] by the explicit inclusion of $g(\theta_1,\chi_1)$ and $g(\theta_2,\chi_2)$. The distribution functions $g(\theta_1)*g(\chi_1)$ and $g(\theta_1,\chi_1)$ for the TIP4P model are shown in Figures 15 and 16. Careful inspection of these figures suggests that $g(\theta_1,\chi_1)$ differs from $g(\theta_1)g(\chi_1)$ quantitatively, which is supported by the two-dimensional Shannon entropy $S^{[\theta_1,\chi_1]}$ differing significantly from the sum of $S^{[\theta_1]}$ and $S^{[\chi_1]}$. For example, for the TIP4P model, the first shell Shannon entropy of $S^{[\theta_1,\chi_1]}$

Table 1. Orientational Shannon Entropies of the Five Water Models[a]

	water models	$S^{[t_1,t_2]}$	$S^{[t_1,\chi_1]}$	$S^{[t_1,\chi_2]}$	$S^{[\chi_1,\chi_2]}$	$S^{[t_1]}$	$S^{[\chi_1]}$
	TIP4P	−1.33	−1.21	−1.15	−1.02	−0.34	−0.29
	SPC	−1.67	−1.28	−1.24	−0.89	−0.50	−0.27
shell 1	TIP3P	−1.65	−1.16	−1.14	−0.74	−0.47	−0.23
	SPC/E	−1.70	−1.32	−1.29	−0.94	−0.51	−0.29
	TIP4P-Ew	−1.44	−1.29	−1.23	−1.05	−0.39	−0.30
	TIP4P	−0.59	−0.44	−0.46	−0.38	−0.10	−0.10
	SPC	−0.69	−0.42	−0.46	−0.30	−0.11	−0.09
shell 2	TIP3P	−0.60	−0.29	−0.34	−0.18	−0.09	−0.06
	SPC/E	−0.71	−0.46	−0.50	−0.33	−0.13	−0.10
	TIP4P-Ew	−0.68	−0.51	−0.53	−0.38	−0.12	−0.12
	TIP4P	−0.010	−0.007	−0.002	−0.003	−0.001	−0.000
	SPC	−0.014	−0.007	−0.005	−0.001	−0.002	−0.000
shell 3	TIP3P	−0.015	−0.003	−0.003	−0.001	−0.002	−0.000
	SPC/E	−0.013	−0.007	−0.005	−0.003	−0.001	−0.000
	TIP4P-Ew	−0.012	−0.007	−0.004	−0.001	−0.001	−0.000

[a] $t = \pi/2(\cos(\theta) + 1)$; all of these entropies are unitless.

Table 2. Comparison of Entropy Results from the NN Method and Cell Theory[a]

	EXP	TIP4P	TIP3P	SPC	SPC/E	TIP4P-Ew
$s^{(2)}_{trans}$		−3.15(3.14[b])	−2.99	−2.99	−3.19	−3.33
$s^{(2)}_{orient}$		−10.52(9.10[b])	−8.58	−10.20	−11.53	−11.76
$s^{(2)}_{ex}$		−13.67(−12.2[b])	−11.57	−13.19	−14.72	−15.09
s_{ex}	−14.05[c]	−14.32[d]	−13.36[d]	−14.01[d]	−14.79[d]	−14.99[d]

[a] Entropies in cal/(mol·K) (eu). [b] Data from Lazaridis.[6] [c] Data from Wagner.[42] [d] Data from Henchman by cell theory.[41]

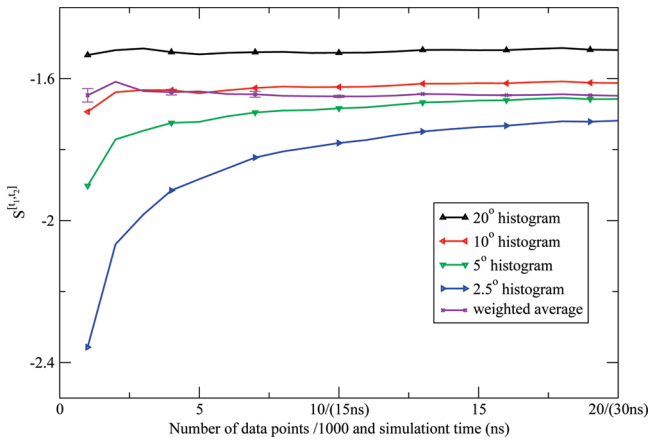


Figure 4. The first shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 1000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using histogram method. The weighted average of the NN estimates and the associated error bar were also depicted.

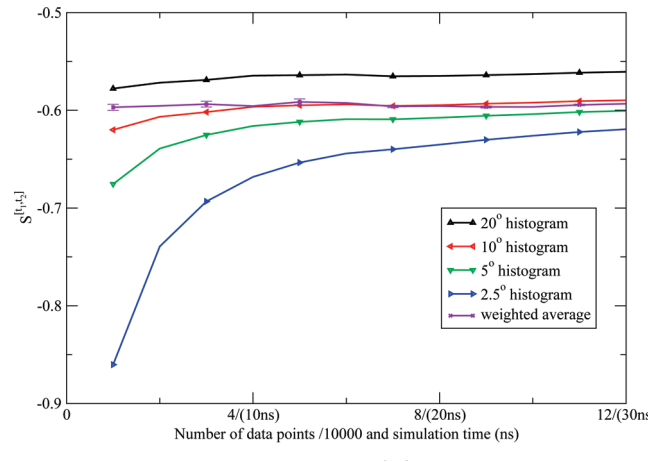


Figure 5. The second shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 10 000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using histogram method. The weighted average of the NN estimates and the associated error bar were also depicted.

is −1.21, while $S^{[\theta_1]}$ is −0.34 and $S^{[\chi_1]}$ is −0.29. This result indicated a non-negligible correlation between χ_1 and θ_1, which suggested that the explicit inclusion of $g(\theta_1,\chi_1)$ and $g(\theta_2,\chi_2)$ in our factorization would lead to greater quantitative precision. This also explains why our excess entropy result for the TIP4P model (−13.67 eu) is about 1.5 eu more negative than the previously reported value (−12.2 eu),[6] which is in better agreement with both the FD estimate of the entropy of the model and the experimental estimate of liquid water.

3.6. Comparison of Free Energy Results. From these simulations, we computed the excess molar energies and excess free energies of the various water models. The results of these calculations for all models studied are listed in Table 3 alongside the relevant literature values. The excess free energies we have obtained here show excellent agreement (within 0.5 kcal/mol uniformly) with the high precision FEP results obtained by Shirts et al.[43] Interestingly, the TIP4P model gives results closest to the experimental quantities.

The SPC/E, TIP4P, and TIP4P-Ew models all give free energy results somewhat closer to the Shirts[43] results than the other models. This may not be accidental. In our calculations, the higher order multiparticle correlation entropies were ignored. There is some literature precedence expecting these higher order contributions to the excess entropy to vanish at the temperature of solid−liquid phase transition.[44,45] Recently, Saija has shown that for the TIP4P

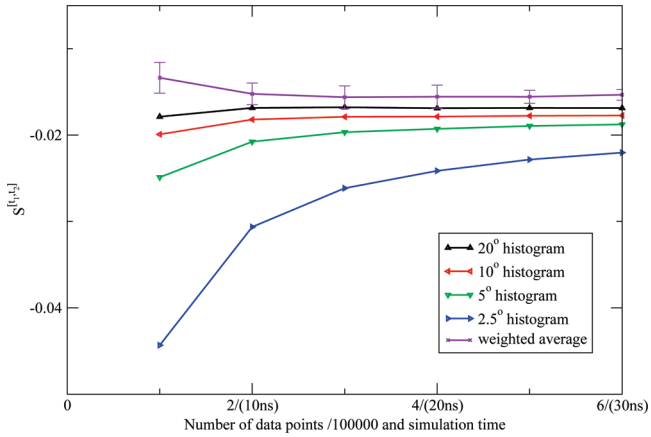


Figure 6. The third shell orientational Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model as a function of the number of data points (labeled on the horizontal axis in front of "/" in units of 100 000) and the corresponding simulation time (labeled on the horizontal axis in parentheses) using histogram method. The weighted average of the NN estimates and the associated error bar were also depicted.

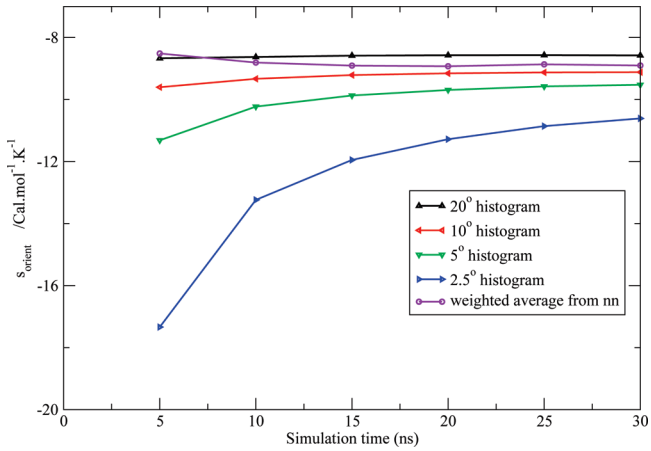


Figure 7. Total orientational excess entropy as a function of simulation time from the NN method and histogram method with different bin width for the TIP3P model.

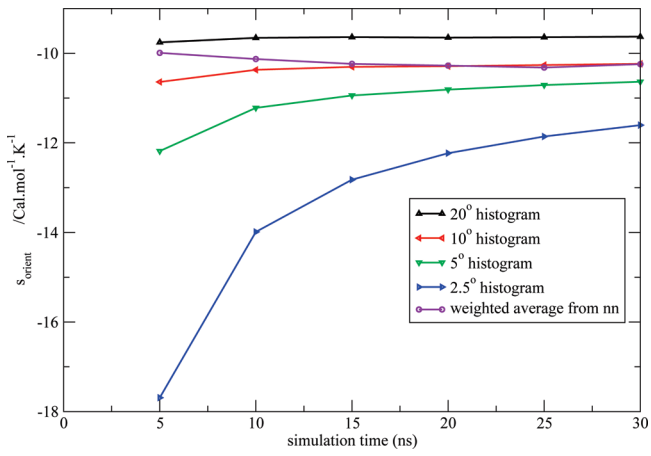


Figure 8. Total orientational excess entropy as a function of simulation time from the NN method and histogram method with different bin width for the SPC model.

model, the temperature of maximum density (TMD) coincides with the temperature where higher order contributions

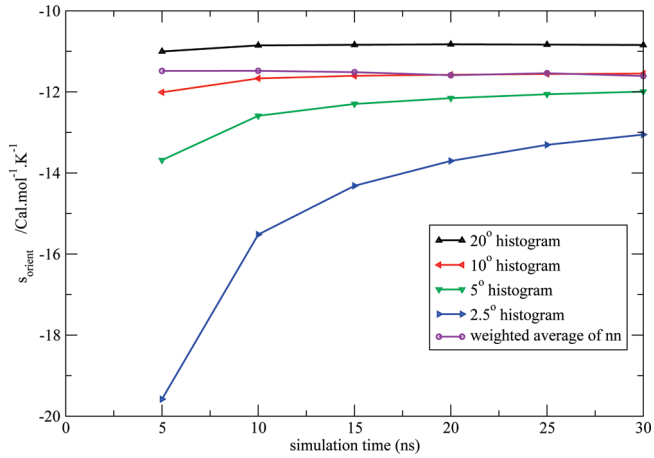


Figure 9. Total orientational excess entropy as a function of simulation time from the NN method and histogram method with different bin width for the SPC/E model.

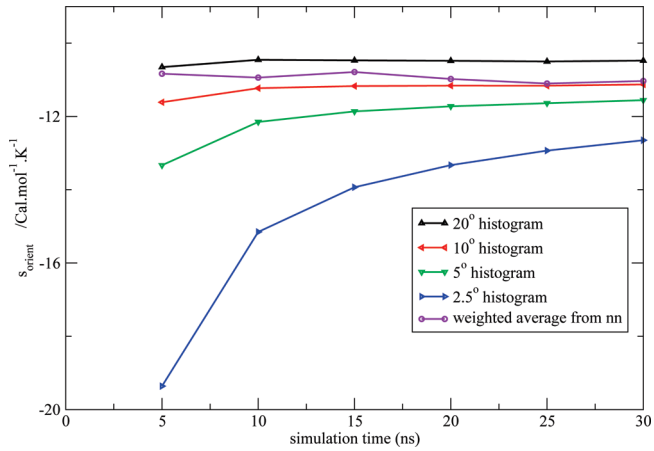


Figure 10. Total orientational excess entropy as a function of simulation time from the NN method and histogram method with different bin width for the TIP4P model.

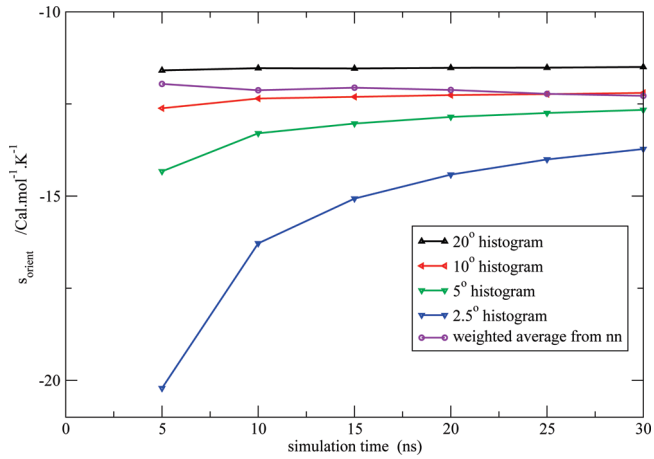


Figure 11. Total orientational excess entropy as a function of simulation time from the NN method and histogram method with different bin width for the TIP4P-Ew model.

to the entropy should vanish.[13] Studies of temperature dependence of the densities of the different water models studied here[46] have shown that the TMD of the TIP4P model occurred at 258 K, the TMD of the SCP/E model occurred at 235 K,[47] the TMD of the TIP4P-Ew model occurred at

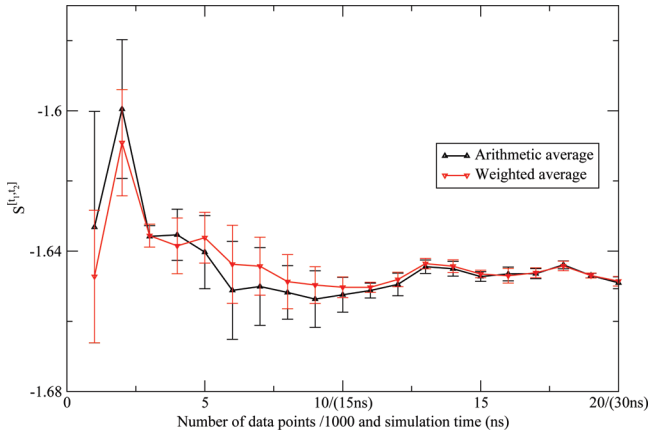


Figure 12. Comparison between the arithmetic average and the weighted average of the NN estimates for the first shell Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model. They are within the error bar of each other.

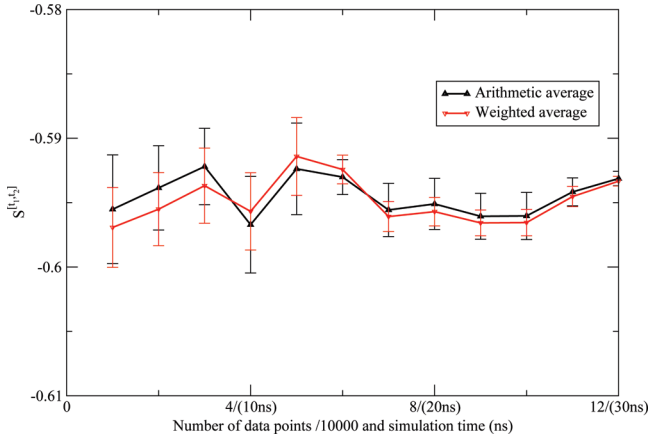


Figure 13. Comparison between the arithmetic average and the weighted average of the NN estimates for the second shell Shannon entropy $S^{[t_1,t_2]}$ for the TIP3P model. They are within the error bar of each other.

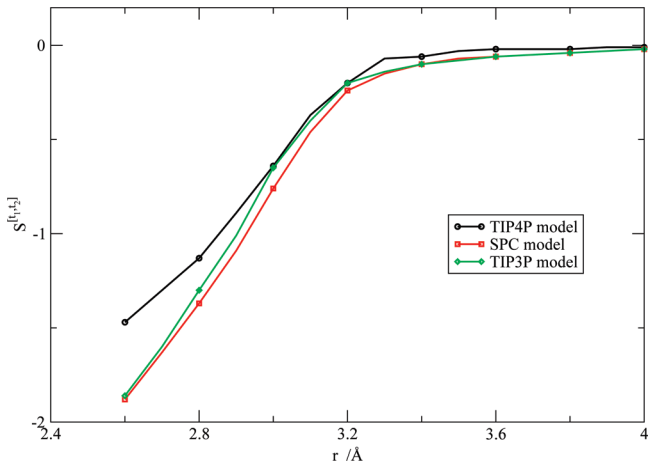


Figure 14. Orientational Shannon entropy $S^{[t_1,t_2]}$ as a function of r for the various water models.

272 K,[40] and the density of the SPC and TIP3P models increases monotonically as temperature decreases in the range [220,370].[46] This indicates, for the TIP3P and SPC models, multiparticle correlation entropy may contribute more to the total entropy than for the other models, which may be why

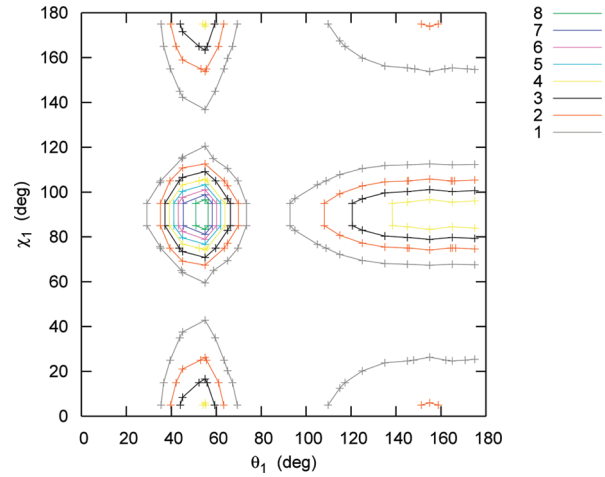


Figure 15. Products of one-dimensional marginal distribution function $g(\theta_1)^*g(\chi_1)$ for the TIP4P model in the first shell.

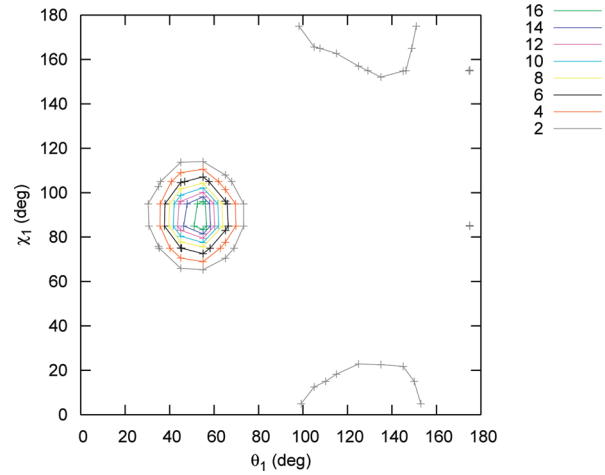


Figure 16. Two-dimensional marginal distribution function $g(\theta_1, \chi_1)$ for the TIP4P model in the first shell.

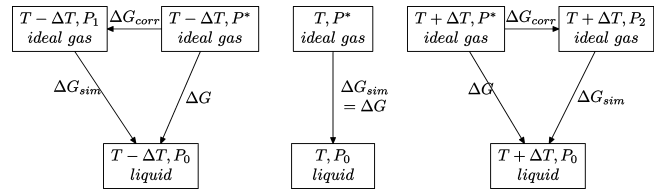


Figure 17. Thermodynamic cycle depicting the constant pressure corrections to ΔG_{sim} at temperatures $T \pm \Delta T$ when computing the slope of ΔG_{sim} with respect to T.

our quantitative accuracy for them is somewhat diminished. However, the molecular detail afforded by this technique in yielding both a value of the entropy and a physical interpretation of its meaning, in terms of the fluid structure implied by the shape of the pair correction function (PCF), gives it a comparative advantage over techniques such as FEP, which will generally only yield a value of the entropy without any additional molecular understanding of the system.

3.7. Entropy Calculation from FD Method. We calculated the excess free energy of water at temperature 298 ± 20 K with the Bennett acceptance ratio[30] method and obtained entropies at 298 K by the FD formula. The results

Table 3. Results for the Energy, Enthalpy, and Entropy of Liquid Water from Various Methods[a]

water models	TIP4P	TIP3P	SPC	SPC/E	TIP4P-Ew
excess energy	−9.85	−9.49	−9.90	−11.08	−10.91
excess enthalpy	−10.43	−10.07	−10.48	−11.66(−10.48[d])	−11.49(−10.45[e])
excess enthalpy[b]	−10.41	−10.09	−10.47	−11.69(−10.51[d])	−11.61(−10.57[e])
excess entropy from NN	−13.67	−11.57	−13.19	−14.72	−15.09
excess entropy[c]	−14.43	−13.39	−14.46	−15.57	−15.53
excess free energy from NN	−6.36	−6.63	−6.55	−7.27(−6.09[d])	−7.00(−5.96[e])
excess free energy[b]	−6.11	−6.10	−6.16	−7.05(−5.87[d])	−6.98(−5.94[e])
excess free energy from exp				−6.33	
excess enthalpy from exp				−10.52	

[a] Energies in kcal/mol, entropies in cal/(mol·K) (eu). [b] Results from Shirts.[43] [c] Results from Shirts[43] by subtracting enthalpy from free energy. [d] Include polarization correction.[39] [e] Include polarization correction.[40]

Table 4. Entropy Results from FD Method and Comparison with Other Methods[a]

water models	TIP4P	TIP3P	SPC	SPC/E	TIP4P-Ew
excess free energy at 278 K	−6.35[b]	−6.21(−6.24[d])	−6.36(−6.39[d])	−7.19(−7.23[d])	
excess free energy at 298 K	−6.03[b]	−5.95	−6.06	−6.89	
excess free energy at 318 K	−5.73[b]	−5.71(−5.69[d])	−5.80(−5.78[d])	−6.66(−6.62[d])	
excess entropy from FD	−15.2[b]	−13.8(±0.8[e])	−15.2(±0.8[e])	−15.3(±0.8[e])	
excess entropy from NN	−13.67	−11.57	−13.19	−14.72	−15.09
excess entropy from FEP[c]	−14.43	−13.39	−14.46	−15.57	−15.53

[a] Energies in kcal/mol, entropies in cal/(mol·K) (eu). [b] Results from Saija.[13] [c] Results from Shirts[43] by subtracting enthalpy from free energy. [d] Results in parentheses include constant pressure correction (Appendix B). [e] Indicates the error associated with the entropy.

are presented in Table 4. The excess entropies computed from the FD method are consistently larger in magnitude than those computed from the NN method, consistent with us neglecting the contributions from the higher order terms of the expansion.

As in the proceeding section, the NN and FD excess entropies of the SPC/E water are in very close agreement; however, the agreement of the NN and FD entropies of the SPC and TIP3P models is much poorer. We again expect the reason for this discrepancy to be due to the TMD of the SPC/E model being close to the range of temperatures treated in this study, while the TMDs of the SPC and TIP3P models fall well outside this range. Thus, the higher order terms of the entropy expansion are expected to make larger contributions to the excess entropies for the SPC and TIP3P models versus the contribution made to the excess entropy of the SPC/E water.

4. Conclusion

Our results indicate that the NN method of computing entropies in the liquid state offers several compelling advantages over the more common histogram approaches, including (1) much faster convergence for a given amount of simulation data; (2) an intuitive error bound for the uncertainty of the calculation without resorting to block averaging or bootstrapping techniques, which may be problematic to apply to estimators of the entropy; and (3) not relying on empirically tuned parameters, such as the histogram bin width, which may bias the results in an unpredictable fashion. We also found that inspection of the limiting behaviours of var ln $f(x)$ may be used to both analyze the convergence of the given calculation and develop the best possible estimate of the entropy given a set of calculated $H_k(n)$. Although we also found that a judicious choice of the histogram bin width may mitigate these advantages, such a choice is difficult to make without prior knowledge of the properties of the limiting distribution, which may not be available when new problems are investigated.

Our alternative factorization of the water−water correlation function, which explicitly included correlations between the angle formed by the water dipole vector and the intermolecular axis with the angle of rotation of the water molecule about its dipole vector, was found to increase the agreement of results obtained by the entropy expansion with those obtained by less approximate methods, such as FEP and the FD benchmark calculations. This result suggests that this contribution should not be ignored in future studies of the excess entropy of liquid water and other fluids.

Acknowledgment. This research was supported by the National Institutes of Health through a grant to R.A.F. (NIH-GM-40526), by the National Science Foundation through a grant to B.J.B. (NSF-CHE-1689) and an NSF Fellowship to R.A., and an allocation of computer time on TeraGrid resources provided by NCSA under NSF auspices.

Appendix A: Determination of Most Proper Weights

Given that $x_1, x_2,...x_n$ are independent variables with the same average u but different variance $v_1, v_2,...v_n$, we may define $\bar{x} = \sum_{i=1}^{n} w_i x_i$, with constraint $\sum_{i=1}^{n} w_i = 1$. We may find the weights w_i such that the variance of $\bar{x}$ is minimized:

$$\mathrm{var}[\bar{x}] = \sum_{i=1}^{n} (w_i)^2 v_i \tag{1}$$

Using Lagrange multipliers, we find:

$$w_i = \frac{\frac{1}{v_i}}{\sum_{i=1}^{n} \frac{1}{v_i}} \tag{2}$$

and

$$\mathrm{var}[\bar{x}] = \frac{1}{\sum_{i=1}^{n} \frac{1}{v_i}} \tag{3}$$

$$E[\sum_{i=1}^{n} w_i(x_i - \bar{x})^2] = E[\sum_{i=1}^{n} w_i((x_i - u) - (\bar{x} - u))^2] \tag{4}$$

$$= E[\sum_{i=1}^{n} w_i((x_i - u)^2 - 2(x_1 - u)(\bar{x} - u) + (\bar{x} - u)^2)] \tag{5}$$

$$= E[\sum_{i=1}^{n} w_i(x_i - u)^2] - 2E[\sum_{i=1}^{n} w_i(x_i - u)(\bar{x} - u)] + E[\sum_{i=1}^{n} w_i(\bar{x} - u)^2] \tag{6}$$

By application of eq 2 and $\sum_i{}^n{}_{=1} w_i = 1$, we find:

$$E[\sum_{i=1}^{n} w_i(x_i - \bar{x})^2] = \frac{n-1}{\sum_{i=1}^{n} \frac{1}{v_i}} \tag{7}$$

Thus, we can approximate the variance of the weighted average by the estimator:

$$V = \frac{1}{n-1} \sum_{i=1}^{n} w_i(x_i - \bar{x})^2 \tag{8}$$

Appendix B: Constant Pressure Correction to ΔG_{sim} for the FD Entropy

In the FEP simulations, we turned on/off the interaction between one distinguished water molecule with the rest of the system at constant temperature T and constant pressure P_0, over the series of several λ windows. The solvation free energy of the distinguished water molecule corresponds to the difference in the chemical potential μ between two phases: (1) the liquid phase and (2) the ideal gas phase with the same temperature and number density as the liquid.[48] For example,

$$\Delta G_{\mathrm{sim}}(T) = -kT \ln \frac{\tilde{\Delta}(\lambda = 1)}{\tilde{\Delta}(\lambda = 0)} = \mu_{\mathrm{l}}(N, P_0, T) - \mu_{\mathrm{g}}(N, P^*, T) \tag{9}$$

where P^* is the pressure of the ideal gas with the same temperature T and number density as the simulated liquid at pressure P_0, and $\tilde{\Delta}$ is the isobaric−isothermal partition function of the system specified by lambda. (For details, please see ref 48.)

The heat capacity of the ideal gas at constant pressure P^* is trivially constant with respect to temperature, and we may well approximate the heat capacity of liquid water to also be constant under constant pressure P_0 over the temperature range studied here. It then follows:

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \tag{10}$$

$$\Delta H(T \pm \Delta T) = \Delta H(T) \pm \Delta C_P \Delta T \tag{11}$$

$$\Delta S(T \pm \Delta T) = \Delta S(T) + \Delta C_P \ln \frac{T \pm \Delta T}{T} \tag{12}$$

$$\Delta S(T) \approx -\frac{\Delta G(T + \Delta T) - \Delta G(T - \Delta T)}{2\Delta T} \tag{13}$$

which are the typical equations of the finite difference method of computing the thermodynamic entropy. In these equations, all of the Δ quantities correspond to the difference of the thermodynamic quantities between the liquid phase at P_0 and the ideal gas phase at P^*.

In similar simulations run at pressure P_0 but temperatures $T \pm \Delta T$, we analogously find

$$\Delta G_{\mathrm{sim}}(T - \Delta T) = \mu_{\mathrm{l}}(N, P_0, T - \Delta T) - \mu_{\mathrm{g}}(N, P_1, T - \Delta T) \tag{14}$$

$$\Delta G_{\mathrm{sim}}(T + \Delta T) = \mu_{\mathrm{l}}(N, P_0, T + \Delta T) - \mu_{\mathrm{g}}(N, P_2, T + \Delta T) \tag{15}$$

where P_1 and P_2 correspond to the ideal gas pressure with the same temperature and number density as the simulated liquids. Note that the ΔG values obtained from simulation differ from those occurring in eq 13 because the reference gas-phase free energies differ, and thus we must explicitly correct for this difference in the reference state. By adding a correction term $\Delta G_{\mathrm{corr}}(T \pm \Delta T)$ to the simulated free energy, we were able to use eq 13 to calculate the entropy at temperature T, where:

$$\Delta G_{\mathrm{corr}}(T - \Delta T) = \mu_{\mathrm{g}}(N, P_1, T - \Delta T) - \mu_{\mathrm{g}}(N, P^*, T - \Delta T) = k(T - \Delta T) \ln \frac{P_1}{P^*} \tag{16}$$

$$\Delta G_{\mathrm{corr}}(T + \Delta T) = \mu_{\mathrm{g}}(N, P_2, T + \Delta T) - \mu_{\mathrm{g}}(N, P^*, T + \Delta T) = k(T + \Delta T) \ln \frac{P_2}{P^*} \tag{17}$$

and

$$\Delta S(T) = -\frac{\Delta G_{\mathrm{sim}}(T + \Delta T) + \Delta G_{\mathrm{corr}}(T + \Delta T) - \Delta G_{\mathrm{sim}}(T - \Delta T) - \Delta G_{\mathrm{corr}}(T - \Delta T)}{2\Delta T} \tag{18}$$

These corrections, although small in magnitude, were systematically of opposite sign at temperatures $T \pm \Delta T$ because the thermal expansion coefficient of liquid water differs from the thermal expansion coefficient of the ideal gas. As a result, failure to apply these corrections will lead to a non-negligible systematical bias in the FD-FEP entropy.

The thermodynamic cycle indicating the whole process, including correction terms, is depicted in Figure 17. Note that in the cycle depicted in Figure 17, we must compute the correction terms at temperatures $T \pm \Delta T$ to compute the slope of ΔG with respect to T, that is, the entropy associated

with the solvation free energy of transferring the water molecule from the gas phase to the liquid phase at temperature *T*.

References

(1) Berne, B. J.; Weeks, J. D.; Zhou, R. *Annu. Rev. Phys. Chem.* **2009**, *60*, 85–103.

(2) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

(3) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 808–813.

(4) Lazaridis, T. *J. Phys. Chem. B* **1998**, *102*, 3531–3541.

(5) Lazaridis, T.; Paulattis, M. E. *J. Phys. Chem.* **1992**, *96*, 3847–3855.

(6) Lazaridis, T.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 4294–4316.

(7) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2006**, *110*, 1464–1475.

(8) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2005**, *109*, 662–670.

(9) Zielkiewicz, J. *J. Phys. Chem. B* **2008**, *112*, 7810–7815.

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

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

(12) Silverstein, K. A. T.; Dill, K. A.; Haymet, A. D. J. *J. Chem. Phys.* **2001**, *114*, 6303–6314.

(13) Saija, F.; Saitta, A. M.; Giaquinta, P. V. *J. Chem. Phys.* **2003**, *119*, 3587–3589.

(14) Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E. *Am. J. Math. Manag. Sci.* **2003**, *23*, 301–322.

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

(16) Raveché, H. J. *J. Chem. Phys.* **1971**, *55*, 2242–2250.

(17) Wallace, D. C. *J. Chem. Phys.* **1987**, *87*, 2282–2284.

(18) Shannon, C. E. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423.

(19) Fisher, I. Z.; Kopeliovich, B. L. *Dokl. Akad. Nauk SSSR* **1960**, *133*, 81–83.

(20) Reiss, H. *J. Stat. Phys.* **1972**, *6*, 39–47.

(21) Singer, A. *J. Chem. Phys.* **2004**, *121*, 3657–3666.

(22) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.

(23) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. *J. Comput. Chem.* **2007**, *28*, 655–668.

(24) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096–3102.

(25) Loftsgaarden, D. O.; Quesenberry, C. P. *Ann. Math. Statist.* **1965**, *36*, 1049–1051.

(26) Arya, S.; Mount, D. M. Approximate nearest neighbor queries in fixed dimensions. *SODA '93: Proceedings of the fourth annual ACM-SIAM Symposium on Discrete Algorithms*; Philadelphia, PA, 1993; pp 271−280.

(27) Freidman, J. H.; Bentley, J. L.; Finkel, R. A. *ACM Trans. Math. Softw.* **1977**, *3*, 209–226.

(28) Smith, D. E.; Haymet, A. D. *J. Chem. Phys.* **1993**, *98*, 6445–6454.

(29) Wan, S. Z.; Stote, R. H.; Karplus, M. *J. Chem. Phys.* **2004**, *121*, 9539–9548.

(30) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(31) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; New York, NY, 2006; p 84.

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

(33) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

(34) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(35) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

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

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

(38) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, 1981; pp 331−342.

(39) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(40) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

(41) Henchman, R. H. *J. Chem. Phys.* **2007**, *126*, 064504.

(42) Wagner, W.; Pruβ, A. *J. Phys. Chem. Ref. Data* **2002**, *31*, 387–478.

(43) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.

(44) Wallace, D. C. *Int. J. Quantum Chem.* **1994**, *52*, 425–435.

(45) Giaquinta, P. V.; Giunta, G. *Physica A* **1992**, *187*, 145–158.

(46) Jorgensen, W. L.; Jenson, C. *J. Comput. Chem.* **1998**, *19*, 1179–1186.

(47) Baez, L. A.; Clancy, P. *J. Chem. Phys.* **1994**, *101*, 9837–9840.

(48) Horn, H. W.; Swope, W. C.; Pitera, J. W. *J. Chem. Phys.* **2005**, *123*, 194504.

CT900078K

JCTC Journal of Chemical Theory and Computation

Dynamically Polarizable Water Potential Based on Multipole Moments Trained by Machine Learning

Chris M. Handley and Paul L. A. Popelier*

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain and School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

Received November 3, 2008

Abstract: It is widely accepted that correctly accounting for polarization within simulations involving water is critical if the structural, dynamic, and thermodynamic properties of such systems are to be accurately reproduced. We propose a novel potential for the water dimer, trimer, tetramer, pentamer, and hexamer that includes polarization *explicitly,* for use in molecular dynamics simulations. Using thousands of dimer, trimer, tetramer, pentamer, and hexamer clusters sampled from a molecular dynamics simulation lacking polarization, we train (artificial) neural networks (NNs) to predict the atomic multipole moments of a central water molecule. The input of the neural nets consists *solely* of the coordinates of the water molecules surrounding the central water. The multipole moments are calculated by the atomic partitioning defined by quantum chemical topology (QCT). This method gives a dynamic multipolar representation of the water electron density without explicit polarizabilities. Instead, the required knowledge is stored in the neural net. Furthermore, there is no need to perform iterative calculations to self-consistency during the simulation nor is there a need include damping terms in order to avoid a polarization catastrophe.

1. Introduction

The existence of life on Earth, and perhaps in the rest of the cosmos, is reliant on the curious and unique properties of water. Liquid water has a number of anomalous properties compared to other similarly sized molecules and displays a great number of solid phases. Understanding water is critical to the understanding of environmental issues,[1–3] green chemistry,[4–7] and biological processes.[8–10] There are a number of excellent reviews on the broad subject of water and its unusual physicochemical features. Water is also an excellent solvent as well as being able to accommodate gases and guest molecules in the form of clathrate hydrates.[11–17]

Water has been the subject of numerous theoretical studies ever since the dawn of computational methods. The first water potential by Bernal and Fowler[18] began what would be a further 40 years of water simulations, with many different potentials designed along the way. However, over time, the focus has shifted from the analysis of small and relatively simple water clusters to larger systems where the number of water molecules is in the hundreds and interacting with other species. As the number and types of atoms involved in the simulation increase, ab initio calculations for these systems become computationally demanding, even exorbitant. Thanks to force fields and water potentials such simulations become tractable, albeit with a loss in accuracy. However, designing an accurate water potential is no easy feat[19] and has been the subject of research for almost 80 years. The design of a potential often becomes a parameter fitting problem aiming at the reproduction of a series of target properties, the number of which has grown as experimental techniques improved. This, along with particular shortcuts and simplifications in the design of a potential, reduces its transferability. Among accurate recent potentials there is the full-dimensional ab initio potential-energy surface for the water dimer of Bowman and co-workers[20] and the first-principles water potential of Bukowski et al.[21]

It is widely recognized that there is a need for polarizability to be included explicitly within water potentials.[19,22–24] It

* Corresponding author e-mail: pla@manchester.ac.uk.

10.1021/ct800468h CCC: $40.75
Published on Web 05/19/2009

has been further suggested that flexibility should not be introduced[25] until polarization is properly modeled. There is also a need for the electron density to be represented more accurately than can be achieved with point charges, better reflecting its anisotropic nature.[26–28] The most popular method is to represent the electron density by multipole moments defined from ab initio wave functions.[29–31]

Recently, artificial learning methods have appeared in potential design. Rather than trying to speculate which functions best model a potential, research groups introduced methods such as neural networks (NNs) to learn the potential from large amounts of ab initio data.[32–35] In this work we present a new method where NNs are trained to learn the relationship between a given water cluster configuration (input) and the multipole moments of an atom in the central water molecule within that configuration (output). After proper training, NNs can predict the multipole moments for an atom (within a water environment) in response to a given water cluster configuration. Hence, the NNs allow the multipole moments to dynamically respond to changes in the local cluster configuration. We developed the above method for dimer, trimer, tetramer, pentamer, and hexamer clusters. In these systems, moments are predicted for the central water molecule based upon the positions of the neighboring water molecules. These dynamic multipole moments combine polarization and charge-transfer effects in a single dynamic correction to the unpolarized Coulomb term. In addition, as we show below, we consider[36] water molecules appearing in water clusters as nonoverlapping. According to QCT (section 2.2), water molecules have finite boundaries and leave no spatial gaps between them. As a result, there is no need to correct for the so-called penetration effect,[37] typically by means of damping functions.

2. Background

2.1. Water Potentials. Water potentials fall into two categories: ab initio potentials and empirical potentials. Ab initio potentials are models where the force field parameters are set to reproduce the potential-energy surface as found by a sample of ab initio calculations.[38–40] Empirical potentials are parametrized to reproduce the bulk phase thermodynamic properties, a well-known example being TIP5P[25] and the more recent TIP4P/2005.[41] Both methods suffer from a lack of transferability. A potential fitted to reproduce the potential-energy surface of small clusters is not an ideal model for bulk conditions. Equally, a potential fitted to reproduce the bulk properties is not the ideal model for a molecule within small cavities and surfaces.

The simplification of using point charges on nuclear positions and at the 'lone pair' positions about the oxygen atom is popular. They are still used in the SPC series of models,[42–44] the TIPS series,[25,45,46] and the model of Nada and van der Eerden[47] who combined TIP4P and TIP5P to create a six-site potential, called NvdE.[48] It is the location and size of these point charges that may be modified to recover the targeted properties of water. However, there are many combinations of charges and water structures (bonds and angles) that will give the correct dipole and quadrupole moments, but this does not mean that any such model will correctly predict further properties.[44] The structure and charge distribution of a water molecule is finely balanced and has an influence on further properties, as expressed by Vega et al. when considering the relative stabilites of ices.[24] This is a view shared by Finney, who finds that classical methods of locating charge, i.e., at 'lone pairs', are not supported by quantum mechanics.[11,12]

It is known that the dipole of water increases from its gas-phase value of 1.85 D[49] to somewhere between 2.3 and 3.1 D[50–54] when moving from the gas phase to the bulk. In fact, water is a very polarizable molecule, able to respond to the electrostatic influence of ions and fields. Water realigns such that it opposes the field. The response to an external field is quickly transmitted through the hydrogen-bonding network. For this reason, liquid water is able to dissolve solids into the component ions. To account for this dipole enhancement, some models have had their charge distribution artificially changed so that the effect is included implicitly, such as in SPC/E[42] and TIP4P.[43] Such models are unreliable for simulations of water in the gas phase, in small cavities, surfaces, or very polar environments.[19,55] It is assumed that a more accurate water model that is transferable to many phases will need to account for polarization correctly.

2.2. Quantum Chemical Topology and Coulomb Interaction. Multipole moments are widely accepted to better represent the electron density of water (and other molecules). Studies by Gresh et al.,[26] Kaminsky and Jensen,[27] and Rasmussen et al.[28] have demonstrated that a multipolar representation of electron density is vital for modeling electrostatic interactions accurately. The multipole moments are coefficients of the series expansion that describes the electrostatic potential generated by an electron density. Multipole moments require more computational resources compared to point charges, even if they are expressed in terms of (irreducible) spherical harmonics as opposed to less compact Cartesian tensors. Multipole moments can be determined from ab initio wave functions by a number of methods, distributed multipole analysis (DMA)[56] being a well-documented and popular one. Multipole moments defined by this method have been successfully employed by Buckingham and Fowler.[57] These moments also turn up in the ASP potential,[29,58] the AMOEBA[31] water model, and the effective fragment potential (EFP) method.[59] Within the "sum of interactions between fragments ab initio" (SIBFA)[60,61] potential, the multipole moments are determined by the partitioning method of Vigné-Maeder and Claverie.[62] A further partitioning method that has grown from SIBFA is the Gaussian electrostatic model (GEM),[63,64] though it relies on density fitting rather than multipole moments.

Within this work, the partitioning of the electron density follows the method of the quantum theory of "atoms in molecules",[65–67] which is part of the quantum chemical topology (QCT) approach. A justification and rationale for the latter name can be found in ref 68. QCT defines topological atoms by the so-called gradient paths in the electron density. Gradient paths originating at infinity follow the direction of steepest ascent in every point of space. They typically (but not necessarily) terminate at nuclei. The three-

dimensional bundle of gradient paths that terminate at a given nucleus defines an atomic volume. A different (two-dimensional) bundle of gradient paths forms an interatomic surface that marks the boundary between two atoms. This bundle terminates at a so-called bond critical point, which lies in between two atoms that share a common boundary. *There are no gaps between topological atoms, and they collectively take up all space.* Atomic multipole moments are calculated by integrating the corresponding property density over the atomic volume. As an integrand of the volume integral, multiplication of the total electron density with regular spherical harmonics gives the required multipole moments.

The electrostatic interaction energy between two atoms is given by eq 1[69]

$$E^{\mathrm{AB}} = \sum_{l_\mathrm{A}=0}^{\infty}\sum_{l_\mathrm{B}=0}^{\infty}\sum_{m_\mathrm{A}=-l_\mathrm{A}}^{l_\mathrm{A}}\sum_{m_\mathrm{B}=-l_\mathrm{B}}^{l_\mathrm{B}} T_{l_\mathrm{A}m_\mathrm{A}l_\mathrm{B}m_\mathrm{B}}(\mathbf{R})Q_{l_\mathrm{A}m_\mathrm{A}}(\Omega_\mathrm{A})Q_{l_\mathrm{B}m_\mathrm{B}}(\Omega_\mathrm{B}) \quad (1)$$

The multipole moments of atom A, $Q_{l_\mathrm{A}m_\mathrm{A}}(\Omega_\mathrm{A})$, and atom B, $Q_{l_\mathrm{B}m_\mathrm{B}}(\Omega_\mathrm{B})$, interact through the tensor $T(\mathbf{R})$. $\mathbf{R}$ is the vector from nucleus A to nucleus B, the origins of the local frames for each atom. Collecting together the terms of eq 1 by their power of $R = \mathbf{R}$, we gather together terms of the same rank, L, defined as $l_\mathrm{A} + l_\mathrm{B} + 1$, where l is the rank of the multipole moment. For example, R^{-3} dependence consists of interactions between two dipole moments ($l_\mathrm{A} = l_\mathrm{B} = 1$) and between a monopole moment ($l = 0$) and a quadrupole moment ($l = 2$). By varying L, the convergence of the multipole expansion can be monitored. Hättig's recurrence formula[70] for the interaction tensor generates expansions up to arbitrarily high rank. The exact interaction energy can be obtained via a six-dimensional integration over the two participating atoms Ω_A and Ω_B

$$E^{\mathrm{AB}} = \int_{\Omega_\mathrm{A}} \mathrm{d}\mathbf{r}_\mathrm{A} \int_{\Omega_\mathrm{B}} \mathrm{d}\mathbf{r}_\mathrm{B} \frac{\rho_{\mathrm{tot}}(\mathbf{r}_\mathrm{A})\rho_{\mathrm{tot}}(\mathbf{r}_\mathrm{B})}{r_{\mathrm{AB}}} \quad (2)$$

where r_{AB} is the distance between two infinitesimally small charge elements and ρ_{tot} is the total charge density (which includes the nuclear charge). Before[71] we made a distinction between the terms "electrostatic" and "Coulomb". The former term is only well defined in the context of (long-range) intermolecular perturbation theory, while the latter applies to the interaction of any charge densities, whether in an intra- or intermolecular context. Since we will sample the electron density from atoms in supermolecules (i.e., water clusters) we are not working in a perturbation context, and hence, the term Coulomb is more appropriate. However, some texts use the two terms interchangeably.

QCT multipole moments are successful in MD simulations of liquid hydrogen fluoride and water[30,72,73] and aqueous solutions of imidazole as well as neat liquid imidazole.[74]

2.3. Polarization. Polarization causes up to a 70% increase in the dipole moment of water, and polarization is often quoted as accounting for ~15% of the total interaction energy[75,76] or as high as 50%.[77] The easiest way to account for polarization is implicitly, fitting the model parameters so that the experimental bulk phase properties are recovered. However, this does not allow for a dynamic anisotropic response of the electron density to an external field and changes in the local chemical environment.

The effect of an electric field upon a molecule can appear in three ways, as outlined by Yu and van Gunsteren.[77] A molecule can respond to an external field by a combination of reorientation, geometrical changes, and electronic redistribution. All models, whether or not they include a geometric or electronic response to polarization, will induce spatial reorientation of a molecule in response to an external field. However, this reorientation will of course be affected by any geometrical and electronic polarization responses that are accounted for by the model. Flexible models with static charge distributions allocated to atoms do account for polarization as a change in geometry in turn changes the molecular electron distribution. However, most water models assume a rigid geometry and concern themselves with the inclusion of electronic polarization.

Polarization can be accounted for explicitly in a number of ways. Three popular methods are (i) polarizable point dipoles,[78] (ii) fluctuating atomic charges,[79] and (iii) attaching a fictitious negative charge[80] to the molecule by a harmonic spring. The danger of the point dipole method is the "polarization catastrophe", where the dipoles respond in such a way that the interaction energy becomes infinite. In order to prevent a polarization catastrophe, where the dipole moments become infinite, a Thole damping function limits the response of the dipole moments.[81–84] The point dipole method appears in the AMOEBA water potential,[31] where polarizable point dipoles are located on atomic centers. Within the SIBFA model, polarizable point dipoles are also situated at off-nuclear positions.[60,85–87] This is analogous to the method in the EFP force fields.[88] The charge-on-spring method refined by MacKerell Jr. and Roux[89] is a simple concept adhered to in the past and more recently in the charge-on-spring class of water potentials by the van Gunsteren group.[90–93] Here polarization is introduced by a negative point charge tethered to the oxygen of the water molecule by a harmonic spring. Finally, the fluctuating charge method allows for the charge at atomic sites to change in response to the external field. This means charge can be redistributed about the molecule or transferred between two molecules. Hence the fluctuating charge method can model both polarization and charge transfer, where there is a partial transfer of charge between the donor and acceptor molecules that are interacting.[94,95] In this view charge transfer appears as a more extreme case of polarization.[96] This method is seen in combination with the TIP4P[79] model and the POL5 model, which is a modification of the TIP5P/ST2 model.[97] Unlike other polarization models, the fluctuating charge approach models polarization and charge transfer together, without additional terms to represent charge transfer.[59,60,85,86] Within the SIBFA model, charge transfer is explicitly represented by further terms.[98]

Recently, Houlding et al. proposed a novel method for incorporating polarization into a simulation of a hydrogen fluoride dimer[99] via dynamic QCT multipole moments. Drawing on large amounts of ab initio data, NNs were trained

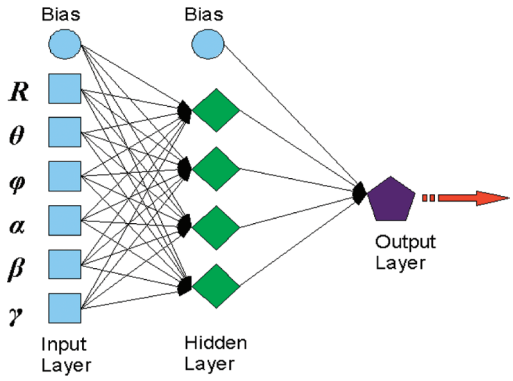


Figure 1. Diagram of a feedforward neural network with one hidden layer. The blue square is an input node, the blue circles are bias nodes, the green diamonds are hidden nodes, and the purple pentagon is an output node. In this work, the output is a multipole moment of a given atom and the inputs are the polar and Euler coordinates of the neighboring water molecules, shown here for the dimer cluster.

to predict the atomic multipole moments of hydrogen fluoride from thousands of configurations of hydrogen fluoride dimers. Following this work, Darley et al. presented a method[100] to tackle *intramolecular* polarization in a similar manner. In that work, polarization was a response to changes in the conformations of *N*-methylacetamide and glycine. In this work we follow the polarization method first introduced by Houlding et al. but expand the method to account for polarization caused by further nearest neighbors.

2.4. Neural Networks. For many years, NNs have been the subject of interest[101–103] as they are useful in discovering mappings between input and output data. In essence they accomplish this by classifying data (possibly large amounts) in arbitrarily high dimensional spaces. A NN is an array of connected nodes, processing units called threshold logic units (TLU), which pass information between themselves. A number of inputs are received by the node, which then sends an output. Each individual input to a node is multiplied by relevant weights, the products subsequently being summed. This sum is then passed through an activation function, whose function value is the final output. Alteration of these weights allows the NN to learn functions and relationships.

The main feature of a NN is the architecture, defined by the number of hidden layers and the number of nodes in the input, output, and hidden layers. Figure 1 shows a NN with inputs for the dimer, a single hidden layer, and an output. There are four hidden nodes and a bias node for both the input and hidden layer. Note that the bias nodes only send signals to the next layer; they do not receive any inputs. The hidden nodes are so named because the user does not have direct access to their outputs and because the hidden neurons must develop their own representation of the inputs.[102] The hidden layer allows the network to learn complex relationships by finding meaningful features from the inputs. The simplest NN is a feedforward network where there is only a single hidden layer of nodes. In a feedforward network, the nodes only pass information to the next layer and not back to a previous layer or to nodes in the same layer. NNs learn the mappings between inputs and outputs from a training set of examples. This *supervised learning* involves the reproduction of a *given* output from the associate input pattern, in order to alter the weights. This is the *backpropagation of errors* method. The process is repeated for every example in the training set before beginning again, with each full pass of the training set called an *epoch*.

Each neuron in layer k sums p inputs x_j from the previous layer j, which are each multiplied by their relevant weight w_{kj}, resulting in activation a_k, as shown in eq 3

$$a_k = \sum_{j=0}^{p} w_{kj} x_j \tag{3}$$

Each weight w_{kj} expresses the relationship between neuron k and neuron j. A weight can be positive or negative for a, respectively, excitatory or inhibitory connection. The output of a neuron must exceed a given threshold, θ, in order to be activating. Typically, the nonlinear sigmoid transfer function determines if a neuron's output is activating, given by eq 4 where ρ defines the shape of the sigmoid and y is the output.

$$y = \sigma(a) = \frac{1}{1 + \exp[-(a - \theta)/\rho]} \tag{4}$$

In this work we are mapping our inputs, the internal coordinates that describe the water clusters, to our outputs, the multipole moments of the atoms in the central water molecule of the cluster.

In order to achieve the optimal NN for the prediction multipole moments the architecture of the NN is modified by varying the number of hidden nodes. NNs predictions can also be improved by altering two training parameters, the learning rate and the momentum.[102] Before training the input data must be standardized, i.e., transformed to dimensionless data that have a mean value of zero and a standard deviation of one. Subsequently, the data are transformed to lie in the interval [0,1] via eq 5

$$x_{i,n} = \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} \tag{5}$$

where x_i are standardized input data, $x_{i,n}$ are normalized input data, and $x_{\min}$ and $x_{\max}$ are minimum and maximum values of the standardized data, respectively.

When training, the NN's performance is determined by the r^2 correlation coefficient, which measures the linear relationship between the predicted output and the desired output, defined in eq 6

$$r^2 = 1 - \left[\frac{\sum_{j=1}^{N} (a_j - b_j)^2}{\sum_{j=1}^{N} \left(a_j - \left(\frac{1}{N} \sum_{j=1}^{N} a_j \right) \right)^2} \right] \tag{6}$$

where a_j is the target output, b_j the predicted output, and N the number of training examples.

A properly trained NN is one that is well generalized. This means that the NN is not overtrained nor overfitted. Overfitting means that the NN suffers from an overly flexible architecture. As a result, it inappropriately absorbs all the

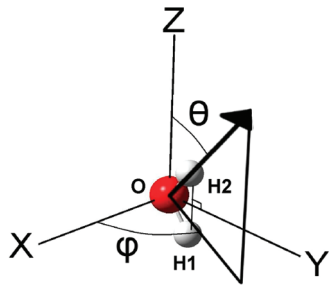


Figure 2. Molecular local frame (MLF) of a central water molecule. The *yz* plane is defined by the HOH plane.

noise in the training set and does not focus on the prime underlying trends in the data. Overtraining refers to when the NN has been trained for too many epochs and offers no predictive ability.

3. Computational Details

Multipole moments allow for an anisotropic description of the electron density of atoms. Multipole moments must be correctly aligned in space relative to each other because the Coulomb interaction between moments is orientation dependent. In our previous work[100] on glycine and *N*-methlyacetamide we introduced an atomic local frame (ALF). In this work we use a molecular local frame (MLF). This means that the atomic multipole moments for a particular water molecule all have the same local frame. This frame, shown in Figure 2, is simply defined by aligning the positive *y* axis along the HOH bisector, with both hydrogen atoms in the positive *y* direction. The *yz* plane is then defined by the HOH plane, with the first hydrogen atom lying at the negative side of the *z* axis.

This orientation convention determines *both* the generation of the training data for the NNs *and* the prediction of the moments by the NNs. The MLF defines the rotation of the moments from the MLF into the global frame. The training data, the multipole moments, are generated with respect to the MLF for every training example. Consequently, the NNs will predict moments with respect to the MLF. This means that the orientation of the MLF is embedded in the training data. The number of cluster configurations (or geometries) needed for the training is determined by the NN's architecture, more specifically, the number of weights. A rule of thumb is to have approximately 10 training examples (i.e., cluster configurations) for every weight being trained. Configurations are taken from molecular dynamics simulations[74] of pure liquid water at ambient conditions performed with nonpolarized (gas phase) QCT multipole moments. For each water molecule in the simulation we find the nearest neighbors that would form the dimer, trimer, tetramer, pentamer, or hexamer clusters. This signifies that we build our clusters in a hierarchical manner about a central water molecule. In other words, the dimer clusters lie within their trimer cluster, the trimer clusters lie within the tetramer clusters, and so forth. This is how approximately 5000 configurations of each cluster size are generated.

The program GAUSSIAN03[104] generated wave functions at the B3LYP/aug-cc-pVTZ level, for each cluster configuration, without geometry optimization. The internal geometry of each water molecule is fixed at the gas-phase-optimized values at the B3LYP/aug-cc-pVTZ level. Due to the modular nature of our method, the electron density may be obtained from wave functions generated at other levels of theory, possibly more advanced, future computing power allowing. For the central water molecule (which lies in the MLF as described above) the program MORPHY[105–107] generated the (atomic) QCT multipole moments for each atom.

The inputs for the NNs are generated from the Cartesian coordinates of the cluster. The Cartesian coordinates are transformed into a set of nonredundant coordinates following the method laid out by Stone.[108] For systems of rigid and nonlinear molecules we have $6(N-1)$ coordinates, where N is the number of molecules. For example, a system of 3 molecules has 12 coordinates, $6(3-1)=12$. With the central molecule at the origin of the MLF and aligned as described above, the position of each neighboring water molecule is described by three polar coordinates and three Euler angles. The polar coordinates are the distance R_{OO} between the central water oxygen atom and the oxygen atom of the neighboring water. The angle θ spans the vector $\boldsymbol{R}_{OO}$ and the z axis, and the angle φ spans the (positive) x axis and the projection of the vector $\boldsymbol{R}_{OO}$ on the xy plane. The three Euler angles, α, β, and γ are measured with respect to the reference water, where the HOH plane lies in the xz plane, with the hydrogen atoms in the negative z direction.

After defining the coordinates of each configuration (for a given cluster size), the data are standardized and normalized for subsequent exploitation by the NNs. For a given atom in the central water molecule, both the coordinates for a given configuration and the multipole moments for the atom in that configuration are transformed to lie between 0 and 1. Training starts operating on these transformed input and output data. Using different network architectures, momenta, and learning rates, NNs are trained for each multipole moment, up to and including the hexadecapole moment ($l=4$), of each atom in the water molecule. To start training, the input data set for a given moment is split into 10 unique validation sets. Each 10% of the data set serves as a validation set in turn. For each assigned validation data the reminder of the data (90%) set is divided up, such that two-thirds are used for training (60%) and the remaining third (30%) for testing in early stopping. In other words, for a given set of training parameters and architecture, we tested for early stopping and validated on 10 different sets.

Training is performed to maximize the training set correlation coefficient r^2. We aim to ensure that the NN is capable of making predictions for examples that were *not* seen in the training set. Hence, we require that the correlation coefficient for the validation set, v^2, is also maximized and close to 1. The statistic v^2 is calculated by the same formula as r^2 (eq 6) but by inserting data of the validation set only. For completeness we also determine the same correlation coefficient for the early stopping set, which we call q^2. The latter statistic is not to be confused with the q^2, the familiar cross-validation correlation coefficient (leave one/many out). However, it is possible that NN is overtrained and offers no predictive ability. To monitor proper NN generalization and

to select the best NN we demand that the ratio r^2/v^2 is close to 1. To certify that the training does not suffer from overtraining or overfitting we monitor the performance of the NN as it is trained. To do this we make use of the early stopping data set. Using the early stopping set we combine two methods for monitoring the performance, which have been described by Prechelt.[109] We test for the loss of generalization of the NN via the generalization loss function, defined in eq 7

$$GL_{(t)} = 100\left(\frac{E_{\mathrm{ea}(t)}}{E_{\mathrm{opt}(t)}} - 1\right) \tag{7}$$

where $E_{\mathrm{opt}(t)}$ is the lowest root-mean-square error (rmse), for the early stopping set, that has ever occurred by epoch t. This corresponds to the best performance ever seen. $E_{\mathrm{ea}(t)}$ is the performance found at a given epoch t. $GL_{(t)}$ is a measure of the loss of generalization at a time t. It is desirable that during the training increasingly better models are generated. If so, the error $E_{ea(t)}$ will always be less than $E_{opt(t)}$, in which case $GL_{(t)}$ is negative. However, if $GL_{(t)}$ is positive then training stops. This testing is initially performed for a given interval of t (here 25 epochs) rather than for every epoch. However, there may be a chance that generalization recovers and improves if training continues. For this reason we test the *progress*. This means that we monitor how many times $GL_{(t)}$ exceeds a specified threshold. If $GL_{(t)}$ exceeds the threshold 10 times then training is ended. When training is signaled to be stopped it returns to the best weights and location on the error surface achieved so far. Generalization loss is then tested, with a lower threshold, at each epoch before training is ended. The initial $GL_{(t)}$ is set to 0.01, which turned out to be adequate for our training purposes. This threshold allows the training to escape local minima in the fitness landscape. Lowering the threshold to 0.005 and progressing epoch by epoch (rather than every 25 epochs) is more appropriate to explore a local minimum. The latter threshold aids in finding the best possible solution without leaping away from it.

This combination of methods described by Prechelt[109] guarantees that training minimizes the errors while avoiding that it is trapped in local minima on the error surface. By training 10 times and testing on 10 different early stopping sets and validation sets we make sure that the training and testing data sets have not introduced a bias and that generalization is maintained without overfitting or overtraining.

The training of single-layer networks occurred with a hidden layer of 4−20 nodes, with momenta of 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, and learning rates of 0.1, 0.25, 0.5, and 0.75.

For *external* testing of the NN's performance a further 1000 configurations were generated. For the dimer clusters, the moments for each of the molecules were generated. These are the *true* moments. Using these moments we can judge the ability of the NNs to predict moments for the water molecules for a given dimer configuration and also the Coloumb interaction energy.

For larger clusters we require a different procedure. For the trimer clusters and larger, 1000 clusters of 50 molecules were generated to act as the test configurations. The arbitrary number 50 were large enough to ensure that each water molecule that is a member of the central clusters we investigate (trimer, tetramer, etc.) can see its own first solvation shell. For the central water, the $n - 1$ nearest neighbors are identified ($n \geq 3$, where n is the total number of molecules in the cluster). From the coordinates of the central molecule and these neighbors the *true* moments are found for the central water molecule. The process is then repeated for each of the $n - 1$ neighbors of the central cluster, where for these neighbors we find their own nearest neighbors and predict moments for these clusters. This ensures that we have the true moments for each of the water molecules based upon their own $n - 1$ nearest neighbors as these are the configurations that are seen when the NNs predict multipole moments for each water molecule in the central cluster of n molecules. NNs are trained to predict moments for a molecule that lies at the center of a cluster of the molecule and its own nearest neighbors. For example, if we were to take the configuration shown in Figure 3a and generate the multipole moments from it, the only water molecule for which these true moments would match the predicted moments is for molecule 2. That is because the true moments and the predicted moments can only be compared if they are in the very same position, that is, where the *molecule is considered at the center of its own nearest neighbors.* It is false to predict moments for molecule 3 based upon the positions of 2 and 1 because molecule 1 is the nearest neighbor of 2 (in the MD simulation from which the cluster is sampled) but not of 3. Instead, the true moments for molecule 3 must be generated based upon *its own nearest neighbors*. Figure 3b shows the actual situation in a small region of the MD simulation from which we sample our test configurations. We predict moments for molecule 1 based upon its own nearest neighbors, 2 and 3. We must then take the true moments for molecule 1 from the wave function of the cluster of molecules 1, 2, and 3. For molecule 2 we predict the moments for this molecule based upon the positions of its own nearest neighbors, namely, 1 and 5. The true moments for molecule 2 are taken from the wave function of the cluster 2, 1, and 5. For molecule 3 its nearest neighbors are in fact 7 and 4. The moments for molecule 3 are predicted based upon the positions of molecules 7 and 4. The true moments are also generated from the wave function of the cluster 3, 4, and 7. Ultimately this procedure must be followed because the *NNs are trained on moments taken from molecules at the center of their own clusters.*

Figure 4 summarizes the sequence of processes of model building and validation. It includes the sampling of the training clusters, the generation of wave functions, the calculation of multipole moments, the training of NNs, the assessment of their prediction performance in terms of the moments themselves, and the Coulomb atom−atom interactions. The upper left corner of Figure 4 starts with the training and test configurations being sampled from the same MD simulation. The wave functions are then calculated for the training configurations ("Gaussian of central cluster") and for the test configurations ("Gaussian of central + neighbour clusters"). From the electron densities correspond-

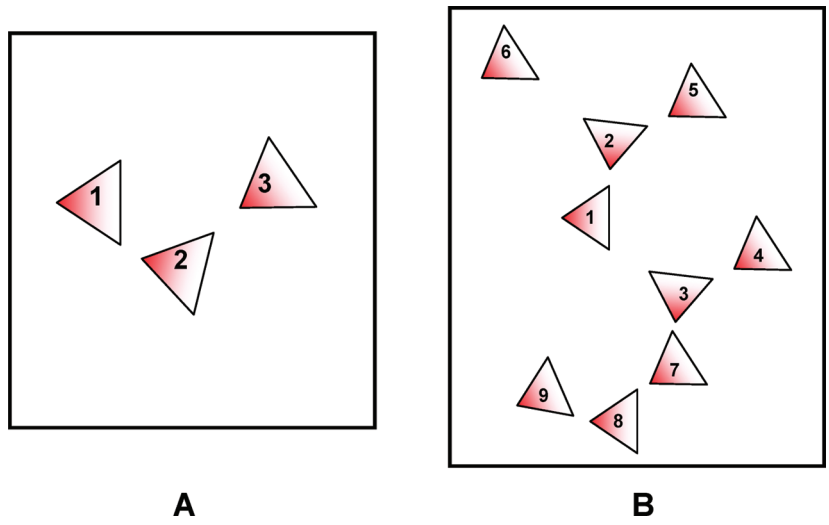


Figure 3. (a) Cartoon of the trimer cluster where molecule 2 is the central molecule of the cluster and molecules 1 and 3 are the nearest neighbors of molecule 2. The red shaded ends of the triangles represent the oxygen atom ends, while the white shaded corners of the triangles represent the hydrogen atom ends. (b) Cartoon of a cluster of 9 water molecules.

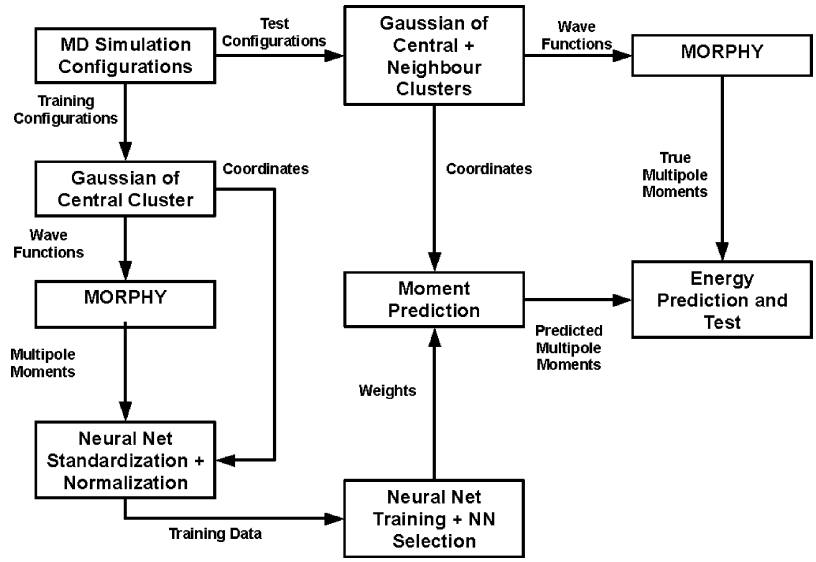


Figure 4. Schematic of the sequence of processes followed to generate and test the models.

ing to all wave functions the multipole moments are computed ("MORPHY"). The training data for the NNs (both input and output) are then prepared prior to actual training ("neural net standardization + normalization"). These training data are then used to train NNs for each moment of each atom. Note that the NNs predict the multipole moments solely based upon the coordinates of the nearest neighbors of a given water molecule. The trained knowledge of the NNs is stored as weights and network architecture. The predicted moments are then confronted with the true moments for a given test configuration (middle right of diagram in Figure 4). The quality of the training is then assessed by the correlation between the predicted and true moments and atom−atom Coulomb interaction energies.

We compared the CPU time required to evaluate all atom−atom interactions by our polarizable multipole model and by the TIP3P potential. On the basis of an average of 1000 water clusters, this overhead is about 61% for dimers and 59% for tetramers.

4. Results and Discussion

Table 1 shows the Pearson correlation coefficient values measured for the training (r^2), early stopping (q^2), and validation (v^2) data sets for the best performing NNs for each moment of the oxygen atom in water dimers. It is clear that generalization is maintained across all the data sets for each of the moments. For each moment the correlation coefficients obtained for the training, early stopping, and validation sets are very similar. Table S1 in the Supporting Information shows the equivalent data for one of the hydrogen atoms, since the data for the other hydrogen are very similar. The correlation coefficients are similar to those found for the oxygen atom. Overall, there is a trend for the performance

Table 1. Statistical Performance of the NN Training of Oxygen in Water Dimers

moment	no. nodes in hidden layer	r^2	q^2	v^2
Q00	10	0.973	0.969	0.973
Q10	9	0.818	0.819	0.816
Q11c	12	0.785	0.761	0.784
Q11s	9	0.916	0.896	0.916
Q20	9	0.992	0.990	0.993
Q21c	10	0.985	0.985	0.986
Q21s	10	0.977	0.976	0.979
Q22c	8	0.984	0.982	0.984
Q22s	11	0.993	0.992	0.992
Q30	11	0.977	0.972	0.977
Q31c	11	0.983	0.982	0.983
Q31s	7	0.949	0.941	0.953
Q32c	11	0.978	0.976	0.978
Q32s	13	0.974	0.971	0.974
Q33c	14	0.980	0.973	0.981
Q33s	11	0.976	0.972	0.978
Q40	12	0.961	0.957	0.961
Q41c	13	0.958	0.951	0.959
Q41s	10	0.990	0.988	0.990
Q42c	15	0.976	0.974	0.976
Q42s	12	0.934	0.924	0.936
Q43c	13	0.648	0.565	0.648
Q43s	14	0.977	0.971	0.977
Q44c	12	0.878	0.873	0.878
Q44s	10	0.862	0.853	0.863

of the NNs to decrease as the rank of the moment increases. However, there are further trends within the set of components of a given moment (i.e., for fixed l in Qlm, where component m varies). In Table 1 we see that for the oxygen atom the y component (Q11s) of the dipole moment is more easily predicted than the two other components, x and z. According to the MLF of water (Figure 2), the x (Q11c) and z (Q10) components of the dipole moment describe the out-of-plane and in-plane deflections of the dipole moment, respectively. These deflections are small. For the dimer cluster test configurations, the average absolute value of the x dipole component is 0.04 and 0.18 D for the z component. These are very small compared to the average value of y dipole component of 2.01 D and with respect to the magnitude of the dipole moment.[54,36] The dipole moment of water is most affected if a nearest neighbor is a hydrogen-bond donor or acceptor. This has a large effect on the magnitude of the dipole moment and thus the magnitude of the y component of the dipole moment. This means that the y component is more easily predicted than the other components because the location of the neighboring molecule has a large influence on it.

The trends seen in the dipole moments in the oxygen atom, Table 1, are not seen in the hydrogen atoms, Table S1 in the Supporting Information. Hydrogen displays almost equal correlation coefficients for all three components of the dipole moment, in contrast to oxygen. The hydrogen atom dipole moments are more dependent upon the precise location of the neighboring water molecule. Hence, the x and z components of a hydrogen dipole moment are easier to predict. For the oxygen atom dipole moment, the main factor determining the dipole moment is which end of the central molecule the nearest neighbor resides at. In other words, the nearest neighbor is either at the oxygen end of the water molecule (negative y semiaxis in Figure 2) or at the hydrogen atom end (positive y semiaxis).

Table 2 shows the decrease in the correlation coefficients for oxygen multipole moments as the cluster size increases from the dimer to the hexamer. As the correlation coefficients decrease, the root mean squared error (rmse) of moment prediction increases, with increasing cluster size. Table S2 of the Supporting Information shows a similar comparison for the hydrogen atoms. For both oxygen and hydrogen, the ability of the NNs to correctly predict the charge and dipole moments diminishes with increasing cluster size. A second feature shared by oxygen and hydrogen is the similarity in r^2 values for their monopole moments. However, the NNs' performance differs dramatically when predicting dipole moments of the hydrogen compared to the oxygen. For example, r^2 values larger than 0.8 still occur for hydrogen dipole components in the hexamer, while r^2 values for oxygen's components can be as low as 0.2. This suggests that the hydrogen atoms are more sensitive to the location of the neighboring molecules. The NNs can cope better with a strong local variation in the dipole moment due to varying positions of the neighboring waters. Conversely, the NNs would be challenged by the more diffuse variation in their dipole moments. In other words, the causal relationship between the position of the water neighbors (input) and oxygen dipole moments (output) is more intricate and buried in the data set. Finally, Table S3 of the Supporting Information shows the correlation coefficients for the NNs predicting the oxygen quadrupole moments with increasing cluster size. It is clear that the ability of the NNs to predict the quadrupole moments is not as adversely affected by the increasing cluster, unlike for the oxygen dipole moments. The majority of the quadrupole moments NNs still have correlation coefficients greater than 0.7 with a fair number even above 0.85. This suggests that the quadrupole moments are more sensitive to the local configuration of the cluster for larger, more homogeneous, bulk-like, clusters.

Table 3 shows the ability of the models to predict the total charge of the central water cluster (in each cluster size). Table 4 shows the ability to predict the (total) dipole moment of the central water (in each cluster). Using the NNs for each model, an in-house code predicted the multipole moments for water molecules and calculated the Coulomb energy from these moments. To calculate the Coulomb interaction between multipole moments of any rank we employ the recursive formula of Hättig.[70] The total dipole moment is described with respect to the MLF. As the cluster size increases, the rise in the average absolute errors for predicting the total charge of the whole cluster and total dipole moment of the central water molecule (within the cluster) is small. We are still able to predict both properties for the larger clusters (tetramer, pentamer, and hexamer) with some accuracy because once we reach these larger cluster sizes the variation of the total charge and total dipole moment is small and less dependent on the local arrangement of the neighboring molecules. This means that in heterogeneous environments, such as small clusters and interfaces, the local arrangement of the neighbors is a critical influence on the total charge and total dipole moment of a water molecule.

Table 2. Statistical Performance of the NN Training of Oxygen in Water with Increasing Cluster Size (rmse in au)

cluster	moment	hidden layer nodes	r^2	r^2 rmse	q^2	q^2 rmse	v^2	v^2 rmse
dimer	Q00	10	0.973	0.040	0.969	0.042	0.973	0.040
	Q10	9	0.818	0.067	0.819	0.068	0.816	0.064
	Q11c	12	0.785	0.070	0.761	0.075	0.784	0.068
	Q11s	9	0.916	0.049	0.896	0.053	0.916	0.048
trimer	Q00	10	0.937	0.041	0.921	0.045	0.937	0.042
	Q10	11	0.621	0.084	0.626	0.088	0.622	0.087
	Q11c	11	0.685	0.078	0.628	0.084	0.683	0.077
	Q11s	16	0.821	0.058	0.764	0.065	0.821	0.059
tetramer	Q00	11	0.885	0.060	0.848	0.068	0.885	0.062
	Q10	10	0.425	0.105	0.284	0.116	0.387	0.108
	Q11c	11	0.501	0.091	0.411	0.098	0.498	0.089
	Q11s	10	0.581	0.086	0.461	0.097	0.518	0.095
pentamer	Q00	22	0.531	0.084	0.433	0.097	0.520	0.089
	Q10	10	0.280	0.107	0.148	0.121	0.278	0.112
	Q11c	12	0.323	0.115	0.171	0.125	0.298	0.121
	Q11s	13	0.271	0.113	0.118	0.122	0.124	0.110
hexamer	Q00	18	0.680	0.080	0.567	0.096	0.666	0.087
	Q10	19	0.314	0.120	0.083	0.136	0.221	0.132
	Q11c	18	0.186	0.140	0.071	0.149	0.159	0.142
	Q11s	17	0.197	0.123	0.085	0.129	0.120	0.129

Table 3. Comparison of the Total Charge Errors (au) of the Central Water Cluster in Each Cluster Size[a]

cluster	average	min	max
dimer	0.0007	−0.0412	0.0464
trimer	–0.0003	–0.0442	0.0527
tetramer	–0.0024	–0.0895	0.0794
pentamer	0.0085	–0.0964	0.1069
hexamer	0.0091	–0.0853	0.1959

[a] The average is taken over all the test configurations for each cluster size.

Table 4. Comparison of Average Absolute Dipole Moment Errors (au) of the Central Water Molecule and the Maximum Absolute Dipole Moment Error for Each Cluster Size[a]

cluster	average abs	max abs
dimer	0.020	0.200
trimer	0.039	0.250
tetramer	0.061	0.312
pentamer	0.072	0.366
hexamer	0.077	0.362

[a] The average is taken over all test set configurations for each cluster size.

The models based upon larger clusters show that the charge and dipole moment of the central water molecule of the cluster is less dependent upon the local arrangement of the nearest neighbors. This is reassuring because the models developed for small clusters are suitable for modeling water in heterogeneous conditions, while our models based on larger clusters are more representative of the polarization response of a water molecule in the bulk phase. This conclusion is supported by the dipole moments predicted for the central water molecule as the cluster size increases.

Figure 5 shows that the models correctly recover the expected dipole moment enhancement of water as we move from the gas phase (1.85 D) to the bulk phase (3.07 D from the work of Bastista et al.[110] and 2.34 D from our previous study[54]). We can also see that the dipole moment of water spans a range of around 0.3 D and that the distributions for each cluster size show signature distribution profiles. The dipole moment for the dimer shows two peaks at 1.93 and 2.11 D, respectively. This is because the dipole moment of a water molecule depends upon whether the nearest neighbor is located at the hydrogen or the oxygen end of the water molecule. This observation highlights that the dipole moment of water is dependent on the hydrogen bonds that the central water molecule is involved in. There is also structure to the distribution for the trimer cluster. In the trimer set there is a main peak at 2.11 D and two lesser peaks at 1.97 and 2.17 D. This can be explained if we consider the arrangements possible of the two neighbors about the central water molecule in the trimer cluster. The two neighbors can both be located at the oxygen end of the water molecule or, alternatively, both at the hydrogen end. The third possibility is that one neighbor is at one end of the central water molecule while the other neighbor resides at the opposite end of this water molecule. As we increase the cluster size further, the distribution of the dipole moments adopts the shape of a bell curve. The peak is now shifted to a higher dipole moment: 2.14 D for the tetramer cluster and 2.27 D for the hexamer. Again, this evolution in the distributions and the shift of the peaks to higher dipole moments suggests that, in building hierarchical water models, we can create a

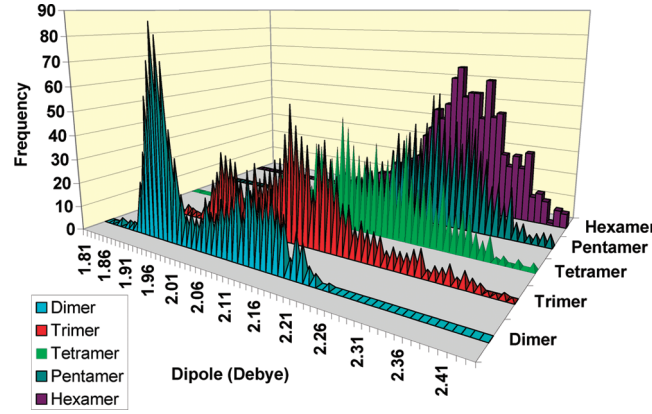


Figure 5. Dipole enhancement effect of water molecules for dimer, trimer, tetramer, pentamer, and hexamer clusters, as predicted by (single layer) NNs.

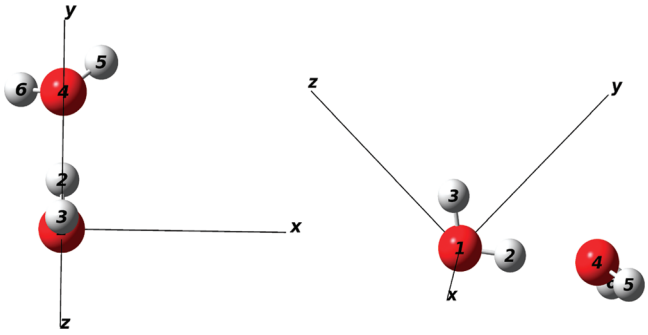


Figure 6. Two perspectives of the dimer configuration selected to test the convergence of intermolecular interactions. This configuration was chosen in view of its similarity to the global minimum in the gas phase. The water molecule H3−O1−H2 is situated at the origin. The global frame (labeled *x*, *y*, and *z*) of the dimer coincides with the MLF of H3−O1−H2, as shown in Figure 2.

range of water models that together are able to describe water moving from the gas phase to the interface and into the bulk phase.

In order to test the validity of the predicted multipole moments the convergence of the multipolar interactions is monitored. A single test configuration sampled from the same MD simulation that is the source of the training data is adequate provided it is common and possessing short-range interaction (i.e., a strong hydrogen bond). This configuration has not been seen by the NNs during training. The configuration, shown in Figure 6, has been selected to resemble the global minimum of the (gas phase) water dimer, differing from it mainly by an asymmetry-inducing tilt of the water molecule H6−O4−H5.

Convergence of multipole moment interaction energies is important if we wish to accurately calculate Coulomb interactions. We dedicated much attention to convergence issues.[69,71,111–116] We can assess the convergence of an interaction by comparing the Coulomb interaction between two atoms for a given rank L to the "exact" Coulomb interaction. The latter is found by a six-dimensional (6D) integration over the volumes of two interacting topological atoms (see eq 2). Since this calculation does not invoke a multipole expansion it can never suffer from convergence problems. We have explicitly shown[100,112] that a given set of multipole moments for two interacting atoms has a particular convergence profile, that is, a plot of the Coulomb interaction energy with respect to rank L. We can use the convergence profile as a way to assess the prediction of the multipole moments. If the NNs were perfect then the predicted multipole moments would be exactly the same as the true moments for the test configuration. In that case the convergence profiles for both sets of multipole moments would be exactly the same. Upon the basis of this principle we can use the difference in convergence behavior as a measure of how well the predicted moments match the true moments. Figure 7 shows the convergence profiles between H2, on one hand, and O4, H5, and H6, on the other. The profiles are calculated by subtracting the exact 6D atom−atom interaction energy from the interaction energy obtained from multipole moments (true or NN predicted). Further convergence profiles for the remaining interactions are shown in the Supporting Information (Figures S1 and S2).

By comparing the convergence profiles from the predicted and true moments for a particular atom−atom interaction it is clear that the main difference in interaction energies is due to the error in the prediction of the atomic monopoles. This can be explained against the apparently contradictory background of their very high value correlation coefficients of prediction. A small relative error for a moment, such as the monopole, translates into a large energy error, compared to the energy errors due to the performance of the other moments. This indicates that correlation coefficients are not the sole judge of the prediction quality. The difference between the convergence profiles for a particular interaction remains almost constant. This means that if we were to further improve the prediction of the interaction energies we must focus on improving the prediction of the monopole moments first. This makes sense since the lower order moments are involved in more moment−moment interactions for a given rank L.

The worst convergence is seen in Figure 7 for the interaction between H2 and O4. This is not surprising considering the short range of this interaction (1.66 Å), while the typical length of a hydrogen bond is about 1.97 Å in liquid water. Compared to the actual magnitude of the atom−atom interactions the errors due to differences in convergence are small (∼0.5−5%) depending upon the type of interaction being computed. Looking at Figures 7 and S1 and S2 in the Supporting Information it is clear that the difference between true and predicted energies is dominated by the charge−charge term. This is because the corresponding energy curves (true and predicted) are roughly parallel with increasing L.

Table S4 of the Supporting Information exhaustively shows how the average $L = 5$ atom−atom interaction energy changes with increasing cluster size up to the pentamer. It lists percentages, which are calculated as 100 E_{true}(A,B) − $E_{\text{predicted}}$(A,B)/E_{true}(A,B), where A and B represent any possible atom pair and "predicted" refers to the NN. The percentages are averaged over all test configurations for a given cluster size. The percentages change very little with increasing cluster size and never rise above 2.5%. This is not unexpected as the ability of the NNs to accurately predict the monopole moments decreases with increasing cluster. The average atom−atom energy only changes in magnitude by up to ∼5% between the dimer and the pentamer (or hexamer).

If we consider the average absolute errors, we see that the errors are smaller for interactions between water molecules that on average become more distantly separated. The average absolute interaction energy errors (for each A,B pair) is on the order of a few kJ/mol, while the average interaction energies are on the order of hundreds of kJ/mol. We do not repeat this table for the nonpolarizable models because the percentages would be embarrassingly high, including for TIP3P. Indeed, in terms of a table such as Table S4 of the Supporting Information, the TIP3P potential performs poorly. We already know that multipolar Coulomb interactions give

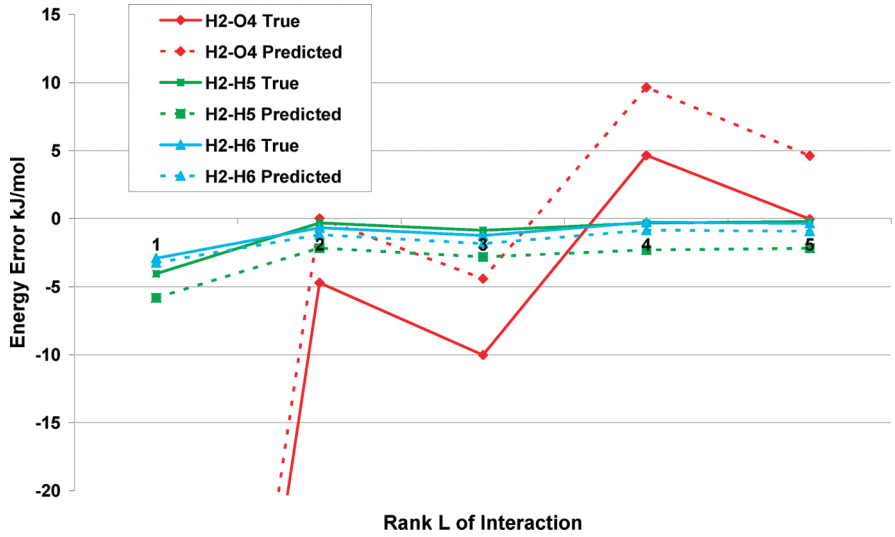


Figure 7. Plot of the convergence of the intermolecular interactions between H2 and O4, H5, and H6, in the water dimer, as the interaction rank *L* increases.

energies close to the energies obtained by 6D integration. Taking the interaction energies calculated using the true multipole moments as *exact energies* we find that TIP3P has average percentage errors of at least 50%, in all interactions, for any cluster size. The same percentage errors occur when unpolarized moments are taken from the monomer wave function of a single water molecule in the gas phase.

Next we compare the total Coulomb interactions, for each cluster size, obtained using the polarized multipole moments, the unpolarized gas phase monomer moments, and the TIP3P point charge model. Figure 8 shows the cumulative plots of the absolute total Coulomb interaction energy errors for the dimer and pentamer clusters, respectively. Further plots for the trimer, tetramer, and hexamer clusters are provided in Figure S3 of the Supporting Information. The "QCT curve" is the cumulative energy error curve found for the static water gas-phase monomer multipole moments. These plots investigate if polarization improves the energy predictions with respect to interaction energies found for the test cluster using the true multipole moments for these clusters. Curiously, "QCT" and TIP3P perform well for the pentamer test configurations. However, inspection of individual atom−atom interaction energies proves that these two models are very wrong and that the accuracy of the total Coulomb interaction is actually a fortuitous cancellation of large errors. The NN-based model is in fact more accurate because the individual atom−atom interactions are also accurate. These curves demonstrate that models cannot just be judged by a single number.

The polarizable model performs better than all other methods except in the case of the pentamer and hexamer clusters. However, from our previous analysis of the atom−atom interaction energies we conclude that for TIP3P and the unpolarized QCT models the interaction energy errors are fortuitous (in the case of the QCT gas-phase monomer moments) or a result of how the model has been fitted. In the case of TIP3P the point charges have been fitted to reproduce the thermodynamic properties of the bulk phase. We know from our previous work[54] that as water cluster size is increased and the number of nearest neighbors about a central water molecule increases, the dipole moment of the water molecule at the center of the cluster increases asymptotically to the bulk phase value. The pentamer cluster size is important for clusters in our hierarchical construction (section 3) as it is at this size that the central water molecule has four nearest neighbors around it. We could consider the pentamer cluster to be the cluster where the central water molecule has obtained its first solvation shell of neighbors. It is also at this cluster size that a large proportion of the dipole enhancement has occurred. It is then possible that the pentamer configurations provide a local environment similar to that of the bulk phase. Since TIP3P is a water model aiming at modeling bulk water it is expected to predict better Coulomb interaction energies than the polarizable model. This in spite of TIP3P's very large errors for the individual atom−atom Coulomb interaction energies. The TIP3P point charges have been parametrized for this type of situation and not for heterogeneous configurations such as the smaller clusters. We are already aware that the point charges can be fitted in a number of ways to reproduce the same properties. However, the differences are clear when we consider the atom−atom interactions. Though TIP3P is a better model for the pentamer clusters, it is right for the wrong reasons. Table 5 summarizes the performance of the models for each cluster size.

Table 5 compares the total Coulomb interaction energies for each cluster size by means of the average absolute errors and the 50th, 90th, and 99th percentile absolute energy errors. This information can be read from Figure 8 and Figure S3 of the Supporting Information. As the cluster size increases, the average absolute total Coulomb interaction error also increases. This is not surprising since the number of atoms for which moments are being predicted is also increasing with cluster size. The latter, in turn, increases the number of prediction errors. However, as the cluster size increases the average total Coulomb interaction energy increases from ∼25 kJ/mol for the dimer to ∼100 kJ/mol for the pentamer clusters. Although the predicted errors are increasing, the

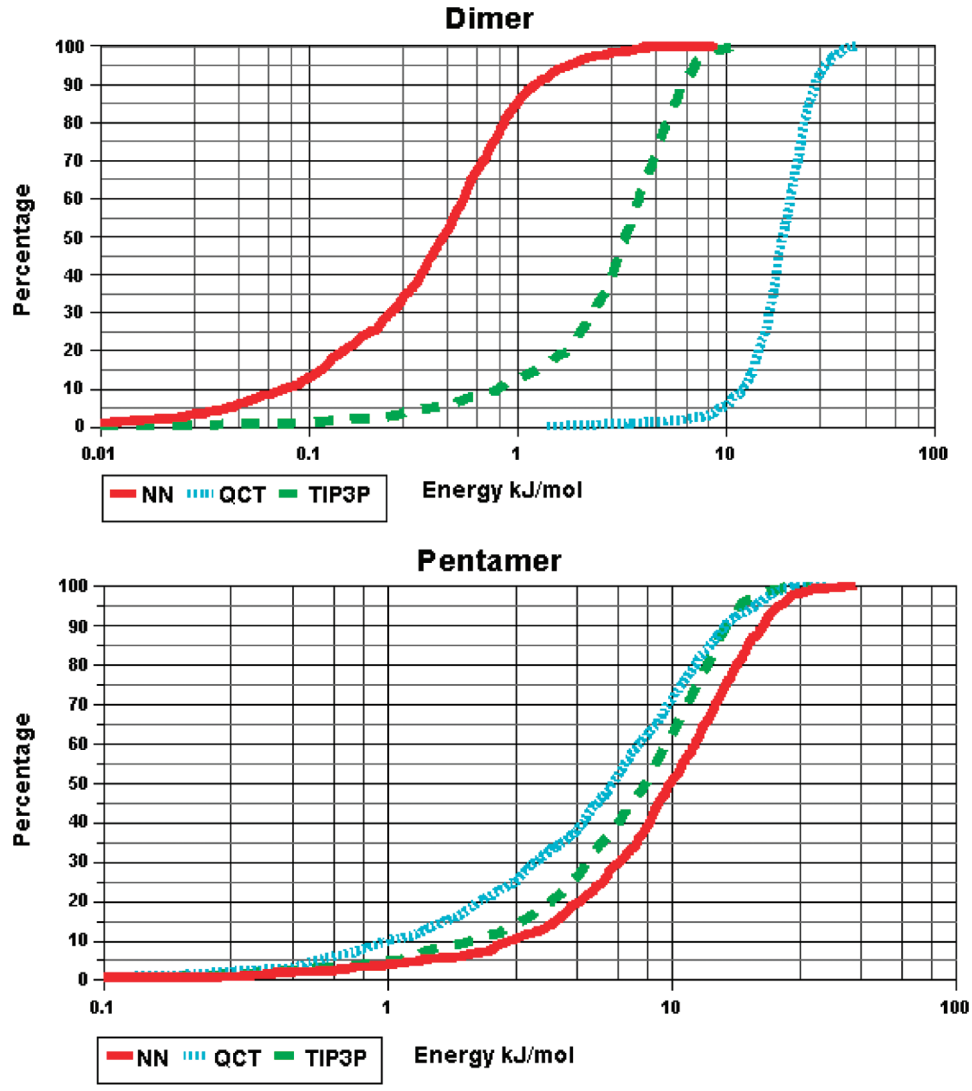


Figure 8. Semilogarithmic cumulative plot of the absolute total Coulomb interaction energy errors for the dimer and pentamer water clusters.

Table 5. Comparison of the Average Absolute Total Coulomb Interaction Energy Errors (in kJ/mol): The 50th, 90th, and 99th Percentile Absolute Total Coulomb Interaction Energy Error

cluster size	model	average	50th percentile	90th percentile	99th percentile
2	NN	0.62	0.44	1.23	3.66
	QCT unpolarized	19.20	18.67	26.86	36.45
	TIP3P	3.52	3.29	6.19	9.61
3	NN	2.59	2.01	5.27	11.44
	QCT unpolarized	10.06	9.62	18.34	24.81
	TIP3P	6.11	4.46	14.07	22.41
4	NN	5.40	4.36	11.37	21.29
	QCT unpolarized	8.86	8.04	16.96	24.55
	TIP3P	6.50	5.11	13.63	25.25
5	NN	11.14	9.89	21.05	30.60
	QCT unpolarized	7.46	6.24	15.43	23.95
	TIP3P	8.61	8.14	15.60	23.12
6	NN	16.37	15.01	27.72	41.89
	QCT unpolarized	8.22	6.41	16.56	29.60
	TIP3P	11.54	10.96	19.27	31.04

percentage errors (calculated as $100\sum_{A,B}^{\text{all pairs}}|E_{\text{true}}(A,B) - E_{\text{predicted}}(A,B)|/E_{\text{true}}(A,B)$) remain fairly constant, at about 10% at worst. More importantly, on average, the NN model for any cluster size is never in error by more than the sum energy of all the possible hydrogen bonds the central water molecule can make (~5 kJ/mol for each hydrogen bond).

At the end of this section we briefly explain how the proposed method avoids the "polarization catastrophe". One

should keep in mind that QCT atoms do not overlap. Instead, the atoms exist as malleable and finite regions in space. Their multipole moments are always well defined, even if the molecules in the clusters come very close to each other. Second, one should remember that the current method does not invoke polarizability tensors. The polarization is implicitly embedded in the trained neural networks. The nets can only produce *finite* multipole moments, and the *effect* of polarization is already fully taken into account when the net generates its multipolar output. On the basis of these important features a polarization catastrophe can never occur.

The proposed polarization approach can be incorporated in molecular dynamics simulations. Inside the MD program DLMULTI (which contains multipolar Ewald summation[117]), we already implemented a subroutine that stores the weights of a trained neural net. The net then predicts, on the fly, the multipole moments of all water molecules in the simulation box, when given their respective environments. In principle, the weights of neural nets obtained by training in one laboratory can be passed on to another lab that performs the simulation. Working within a rigid body framework, forces and torques can be obtained by differentiating eq 1. Note that the differentiation product rule then yields three terms, but only the term with the differentiation of the interaction tensor T is currently included. The nonlinear behavior and anisotropy of polarization is in principle captured by the net and made available to the simulation engine.

5. Conclusions

We proposed a novel, polarizable multipolar water potential that uses neural networks (NNs) to predict atomic multipole moments for any given water cluster configuration, in principle. This method eliminates the need to perform iterations during an MD simulation, unlike for many polarizable models. Second, this method allows for the Coulomb interaction to include polarization and charge transfer, treated on a par, as part of a dynamic Coulomb interaction term. The new water models have been critically analyzed at all stages during the prediction of cluster Coulomb interaction energies in order to assess the ability of the NNs to predict multipole moments on the atoms of a water molecule at the center of a particular cluster. The performance of NNs diminishes with increasing cluster size, which is related to two problems. The first is that the space in which the NNs are trying to fit functions is increasing in dimensionality. The second is that, as the cluster size increases, certain multipole moments become less sensitive to the location of the nearest neighbor water molecules about the central water molecule.

In summary, we critically assessed the NNs' ability to predict the moments of the atom. Additionally, we can further test the NNs by using the predicted multipole moments to compute the charge of the atoms, the net charge of the central water molecule, the dipole moment of the central water molecule, and the atom−atom and total cluster Coulomb interaction energies. We conclude that the NN-based model allows for a more complete description of the Coulomb interactions between water molecules, compared to the nonpolarizable QCT model and the TIP3P water potential. Using the gas-phase monomer moments as a reference, we find that, on average, the polarization of the multipole moments accounts for 50% of the atom−atom Coulomb interaction energies. We showed that, due to the fitting method behind TIP3P, it is able to predict total Coulomb interaction energies that are more accurate than the polarizable model. However, this accuracy is due to a fortuitous cancelation of very inaccurate atom−atom interaction energies. This suggests that other empirical point charge models, fitted in a fashion similar to TIP3P, may again yield accurate total interaction energies for large clusters of water molecules but not for the correct reason.

Fixed charge density models do not accurately predict Coulomb interaction energies for clusters that are representative of the type of local configurations that a water molecule would experience in heterogeneous conditions, such at the water−air interface. Our pentamer model is probably appropriate for the simulation of water in the bulk phase, while the models developed based on smaller water clusters are suited to predicting the Coulomb interaction energies between water molecules in the gas phase and at an interface. Using this hierarchy of models, it would be possible to perform simulations of the ice surface melting where the Coulomb interactions are represented accurately using dynamic polarizable multipole moments.

We are already exploring more elaborate NNs to predict the multipole moments. Two hidden layers in a NN introduce more flexibility to fit a function to the training data. This added flexibility, to be reported on in a future article, has already improved the accuracy of the prediction of the monopole moments of oxygen and improved the Coulomb interaction energies for the dimer, trimer, and tetramer clusters. We are also exploring the use of other statistical learning machines called radial basis function networks, and Kriging models, which already show a better ability to predict accurate polarizable multipole moments compared to the NNs.

Acknowledgment. We thank the EPSRC for financial support through grant EP/C015231 and Dr. G. I. Hawe for valuable comments.

Supporting Information Available: Convergence of the intermolecular interactions between O1 and O4, H5, and H6, in the water dimer, as the interaction rank L increases; convergence of the intermolecular interactions between H3 and O4, H5, and H6, in the water dimer, as the interaction rank L increases; cumulative plot of the absolute total electrostatic interaction energy for the trimer, tetramer, and hexamer systems; statistical performance of the NN training of a representative hydrogen atom in water dimers; statistical performance of the NN training of hydrogen in water with increasing cluster size (rmse in au); statistical performance of the NN training of oxygen in water with increasing cluster size (rmse in au); comparison of atom−atom total interaction energy errors (percentage) with increasing cluster size. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dash, J. G.; Fu, H.; Wettlaufer, J. S. *Rep. Prog. Phys.* **1995**, *58*, 115.

(2) Hervig, M.; Thompson, R. E.; McHugh, M.; Gordley, L. L.; Russell, J. M., III.; Summers, M. E. *Geophys. Res. Lett.* **2001**, *28*, 971.

(3) Prenni, A. J.; DeMott, P. J.; Kreidenweis, S. M. *Atmos. Environ.* **2003**, *37*, 4243.

(4) Savage, P. E. *Chem. Rev.* **1999**, *99*, 603.

(5) Dabiri, M.; Baghbanzadeh, M.; Nikcheh, M. S.; Arzroomchilar, E. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 436.

(6) Gao, R.; Dai, W.-L.; Le, Y.; Yang, X.; Cao, Y.; Li, H.; Fan, K. *Green Chem.* **2007**, *9*, 878.

(7) Chen, L.; Li, C.-J. *Adv. Synth. Catal.* **2006**, *348*, 1459.

(8) Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. *Proc. Natl. Acad. Sci.* **1994**, *91*, 1089.

(9) Okada, T.; Fujiyoshi, Y.; Silow, M.; Navarro, J.; Landau, E. M.; Shichida, Y. *Proc. Natl. Acad. Sci.* **2002**, *99*, 5982.

(10) Killian, J. A.; von Heijne, G. *Trends Biochem. Sci.* **2000**, *25*, 429.

(11) Finney, J. L. *J. Mol. Liq.* **2001**, *90*, 303.

(12) Finney, J. L. *Philos. Trans. R. Soc. London, Ser. B: Biol. Sci.* **2004**, *359*, 1145.

(13) Ludwig, R. *Angew. Chem., Int. Ed* **2001**, *40*, 1808.

(14) Robinson, G. W.; Zhu, S.-B.; Singh, S.; Evans, M. W. *Water in Biology, Chemistry and Physics*; World Scientific Publishing: Singapore, 1996.

(15) Franks, F. *Water: a comprehensive treatise*; Plenum Press: New York, 1972.

(16) Franks, F. *Water: a matrix of life*; Royal Society of Chemistry: Cambridge, 2000.

(17) Stillinger, F. H. *Science* **1980**, *209*, 451.

(18) Bernal, J. D.; Fowler, R. H. *J. Chem. Phys.* **1933**, *1*, 515.

(19) Guillot, B. *J. Mol. Liquids* **2002**, *101*, 219.

(20) Huang, X.; Braams, B. J.; Bowman, J. M. *J. Phys. Chem. A* **2006**, *110*, 445.

(21) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *Science* **2007**, *315*, 1249.

(22) Paricaud, P.; Predota, M.; Chialvo, A. A.; Cummings, P. T. *J. Chem. Phys.* **2005**, *122*, 244511.

(23) Chen, B.; Xing, J.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 2391.

(24) Vega, C.; Sanz, E.; Abascal, J. L. F. *J. Chem. Phys.* **2005**, *122*, 114507.

(25) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.

(26) Gresh, N.; Kafafi, S. A.; Truchon, J.-F.; Salahub, D. R. *J. Comput. Chem.* **2004**, *25*, 823.

(27) Kaminsky, J.; Jensen, F. *J. Chem. Theor. Comput.* **2007**, *3*, 1774.

(28) Rasmussen, T. D.; Ren, P.; Ponder, J. W.; Jensen, F. *Int. J. Quantum Chem.* **2006**, *107*, 1390.

(29) Millot, C.; Stone, A. J. *Mol. Phys.* **1992**, *77*, 439.

(30) Liem, S.; Popelier, P. L. A. *J. Chem. Phys.* **2003**, *119*, 4560.

(31) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

(32) Prudente, F. V.; Acioli, P. H.; Soares Neto, J. J. *J. Chem. Phys.* **1998**, *109*, 8801.

(33) Gassner, H.; Probst, M.; Lauenstein, A.; Hermansson, K. *J. Phys. Chem. A* **1998**, *102*, 4596.

(34) No, K. T.; Chang, B. H.; Kim, S. Y.; Jhon, M. S.; Scheraga, H. A. *Chem. Phys. Lett.* **1997**, *271*, 152.

(35) Cho, K.-H.; No, K. T.; Scheraga, H. A. *J. Mol. Struct.* **2002**, *641*, 77.

(36) Devereux, M.; Popelier, P. L. A. *J. Phys. Chem. A* **2007**, *111*, 1536.

(37) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300.

(38) Matsuoka, O.; Clementi, E.; Yoshimine, M. *J. Chem. Phys.* **1975**, *64*, 1351.

(39) Niesar, U.; Corongiu, G.; Clementi, E.; Kneller, G. R.; Bhattacharya, D. K. *J. Phys. Chem.* **1990**, *94*, 7949.

(40) Lie, G. C.; Clementi, E. *Phys. Rev. A* **1985**, *33*, 2679.

(41) Vega, C.; Abascal, J. L. F.; Conde, M. M.; Aragones, J. L. *Faraday Discuss.* **2008**, *141*, 1.

(42) Berendsen, H.; Postma, J.; Van Gunsteren, W.; Hermans, J. *Interaction models for water in relation to protein hydration*; Reidel: The Netherlands, 1981.

(43) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.

(44) Glättli, A.; van Gunsteren, W. F.; Daura, X. *J. Chem. Phys.* **2002**, *116*, 9811.

(45) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335.

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

(47) Nada, H.; van der Eerden, J. P. J. M. *J. Chem. Phys.* **2003**, *118*, 7401.

(48) Bishop, C. L.; Pan, D.; Liu, L. M.; Tribello, G. A.; Michaelides, A.; Wang, E. G.; Slater, B. *Faraday Discuss.* **2008**, *141*, 1.

(49) Clough, S. A.; Beers, Y.; Klein, G. P.; Rothman, L. S. *J. Chem. Phys.* **1973**, *59*, 2254.

(50) Coulson, C. A.; Eisenberg, D. *Proc. R. Soc. London, Ser. A* **1966**, *291*, 445.

(51) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308.

(52) Gregory, J. K.; Clary, D. C.; Liu, K.; Brown, M. G.; Saykally, R. J. *Science* **1997**, *275*, 814.

(53) Gubskaya, A. V.; Kusalik, P. G. *J. Chem. Phys.* **2002**, *117*, 5290.

(54) Handley, C. M.; Popelier, P. L. A. *Synth. React. Inorg. Met.-Org. Nano-Met. Chem.* **2008**, *38*, 91.

(55) Kollman, P. A. *Acc. Chem. Res.* **1996**, *29*, 461.

(56) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon: Oxford, 1996.

(57) Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.* **1985**, *63*, 2018.

(58) Millot, C.; Soetens, J.-C.; Martins-Costa, M. T. C.; Hodges, M. P.; Stone, A. J. *J. Phys. Chem.* **1998**, *102*, 754.

(59) Gordon, M. S.; Slipchenko, L.; Li, H.; Jensen, J. H. *Annu. Rep. Comp. Chem.* **2007**, *3*, 177.

(60) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856.

(61) Piquemal, J.-P.; Williams-Hubbard, B.; Fey, N.; Deeth, R.; Gresh, N.; Giessner-Prettre, C. *J. Comput. Chem.* **2003**, *24*, 1963.

(62) Vigne-Maeder, F.; Claverie, P. *J. Chem. Phys.* **1988**, *88*, 4934.

(63) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.

(64) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Comput. Chem.* **2007**, *3*, 1960.

(65) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, 1990.

(66) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education: London, 2000.

(67) Matta, C. F.; Boyd, R. J. *The Quantum Theory of Atoms in Molecules*; Wiley-VCH: Weinheim, Germany, 2007.

(68) Popelier, P. L. A.; Aicken, F. M. *ChemPhysChem* **2003**, *4*, 824.

(69) Popelier, P. L. A.; Joubert, L.; Kosov, D. S. *J. Phys. Chem. A* **2001**, *105*, 8254.

(70) Haettig, C. *Chem. Phys. Lett.* **1996**, *260*, 341.

(71) Popelier, P. L. A.; Kosov, D. S. *J. Chem. Phys.* **2001**, *114*, 6539.

(72) Liem, S.; Popelier, P. L. A.; Leslie, M. *Int. J. Quantum Chem.* **2004**, *99*, 685.

(73) Liem, S. Y.; Popelier, P. L. A. *J. Chem. Theory Comput.* **2008**, *3*, 353.

(74) Shaik, M. S. PhD thesis, Design of a Multipolar Potential leading to a Description of the Self-Assembly of Imidazole in Aqueous Solution, The University of Manchester, 2008.

(75) Friesner, R. A. Modeling Polarization in Proteins and Protein−Ligand Complexes: Methods and Preliminary Results. In *Advances in Protein Chemistry*; Baldwin, R.; Baker, D., Eds.; Academic Press: New York, 2005; Vol. 72, p 79.

(76) Hodges, M.; Stone, A. J.; Cabaleiro Lago, E. *J. Phys. Chem.* **1998**, *102*, 2455.

(77) Yu, H.; van Gunsteren, W. F. *Comp. Phys. Commun.* **2005**, *172*, 69.

(78) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237.

(79) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.

(80) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556.

(81) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.

(82) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Dehez, F.; Ángyán, J. G.; Orozco, M.; Chipot, C.; Luque, F. J. *J. Chem. Theory Comput.* **2007**, *3*, 1901.

(83) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.

(84) Gao, J.; Habibollazadeh, D.; Shao, L. *J. Phys. Chem.* **1995**, *99*, 16460.

(85) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353.

(86) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem.* **2007**, *111*, 8170.

(87) Ledecq, M.; Lebon, F.; Durant, F.; Giessner-Prettre, C.; Marquez, A.; Gresh, N. *J. Phys. Chem. B* **2003**, *107*, 10640.

(88) Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, 105.

(89) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587.

(90) Yu, H.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *121*, 9549.

(91) Yu, H.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221.

(92) Yang, M.; Senet, P.; Alsenoy, C. V. *Int. J. Quantum Chem.* **2005**, *101*, 535.

(93) Yu, H.; Geerke, D. P.; Liu, H.; van Gunsteren, W. F. *J. Comput. Chem.* **2006**, *27*, 1494.

(94) Hagberg, D.; Karlstrom, G.; Roos, B. O.; Gagliardi, L. *J. Am. Chem. Soc.* **2005**, *127*, 14250.

(95) Hemmingsen, L.; Amara, P.; Ansoborlo, E.; Field, M. J. *J. Phys. Chem. A* **2000**, *104*, 4095.

(96) Chen, J.; Martínez, T. J. *Chem. Phys. Lett.* **2007**, *438*, 315.

(97) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 5.

(98) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1982**, *22*, 199.

(99) Houlding, S.; Liem, S. Y.; Popelier, P. L. A. *Int. J. Quantum Chem.* **2007**, *107*, 2817.

(100) Darley, M. G.; Handley, C. M.; Popelier, P. L. A. *J. Chem. Theory Comput.* **2008**, *4*, 1435.

(101) Vapnik, V. N. *Statistical Learning Theory*; John Wiley: New York, 1998.

(102) Gurney, K. *An Introduction to Neural Networks*; Taylor and Francis: London, 1997.

(103) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice-Hall, New Jersey, 1999.

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

(105) Popelier, P. L. A.; Bone, R. G. A. *MORPHY98*, UMIST: Manchester, 1998.

(106) Popelier, P. L. A. *Mol. Phys.* **1996**, *87*, 1169.

(107) Popelier, P. L. A. *Chem.Phys.Lett.* **1994**, *228*, 160.

(108) Stone, A. J. *The Theory of Intermolecular Forces*, 1st ed; Clarendon Press:, 1996; Vol. 32.

(109) Prechelt, L. *Neural Networks* **1998**, *11*, 761.

(110) Batista, E. R.; Xantheas, S. S.; Jónsson, H. *J. Chem. Phys.* **1998**, *109*, 4546.

(111) Rafat, M.; Popelier, P. L. A. *J. Chem. Phys.* **2005**, *123*, 204103.

(112) Rafat, M.; Popelier, P. L. A. *J. Chem. Phys.* **2006**, *124*, 144102.

(113) Rafat, M.; Popelier, P. L. A. *J. Comput. Chem.* **2007**, *28*, 832.

(114) Popelier, P. L. A.; Rafat, M. *Chem. Phys. Lett.* **2003**, *376*, 148.

(115) Kosov, D. S.; Popelier, P. L. A. *J. Chem. Phys.* **2000**, *113*, 3969.

(116) Kosov, D. S.; Popelier, P. L. A. *J. Phys. Chem. A* **2000**, *104*, 7339.

(117) Leslie, M. *Mol. Phys.* **2008**, *106*, 1567.

CT800468H

JCTC Journal of Chemical Theory and Computation

Optimization of Capping Potentials for Spectroscopic Parameters in Hybrid Quantum Mechanical/Mechanical Modeling Calculations

Sittipong Komin† and Daniel Sebastiani*

Max-Planck-Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany

Received December 2, 2008

Abstract: We present a capping scheme for hybrid calculations which is designed for a systematic optimization to reproduce the molecular structure, frontier bond potential, and spectroscopic properties for the quantum subsystem. Our technique is capable of reducing the perturbations of the electronic structure which are normally caused by conventional link atoms between quantum and classical regions. Specifically, we propose analytic effective core potentials with a small set of adjustable parameters, which are optimized to reproduce the full-quantum-mechanical (full-QM) properties in the direct environment of the bond cleavage. The capping potentials are conceptually simple and easy to employ in most instances without significant code modifications. They do not require any further external geometry constraints and yield also reasonable results for the potential energy surface. We benchmark these potentials for a series of chemically and biologically relevant molecules calculating NMR chemical shifts, protonation energies, and optimized geometries. Our optimized QM/mechanical modeling (MM) potentials are another step toward a realistic first-principles prediction of spectroscopic parameters in complex chemical environments using hybrid QM/MM calculations.

1. Introduction

The determination of the detailed microscopic structure and dynamics of complex supramolecular systems is still a challenge for modern physics and chemistry. The interplay of intramolecular and intermolecular interactions is crucial for a broad range of chemical, biological, and physical processes that occur in nature.[1–4] To obtain structural data of supramolecular systems, the combination of spectroscopic experiments with advanced theoretical predictions and computer simulations is becoming increasingly popular, because this combination often yields a predictive power above the sum of the individual techniques.[5–7]

With the recent advances in computational methodology as well as computer hardware, the first-principles prediction of such noncovalent effects on the structure and experimentally observable spectra has come into reach for many systems of technological and fundamental scientific interest.[8–10] Several methods exist to incorporate the influence of the chemical environment into such electronic structure calculations. The explicit consideration of a large number of neighboring molecules is in principle most accurate, but computationally very demanding and thus only applicable in simple cases.[11–14]

Alternatively, one can resort to embedding schemes, which can treat the environment at various levels of approximation. In this context, a hybrid method is often adopted which splits the total system into a smaller part, which is treated quantum-mechanically (QM) using electronic structure methods, and the remaining part, which is described via parametrized potentials (MM).[15–17] One of the difficulties of such a hybrid quantum mechanical/mechanical modeling (QM/MM) approach is the transition region between the two different parts. Often, chemical bonds are "broken"; i.e., one of the atoms involved in the covalent bond is in the quantum (QM) part, the other in the classical (MM) one. This situation is sketched

* Corresponding author phone: 49-6131-379-260; fax: 49-6131-379-100; e-mail: sebastia@mpip-mainz.mpg.de.

† Department of Physics, Faculty of Science, Ubonratchathanee University, Ubonratchathanee, Thailand.

10.1021/ct800525u CCC: $40.75
Published on Web 05/04/2009

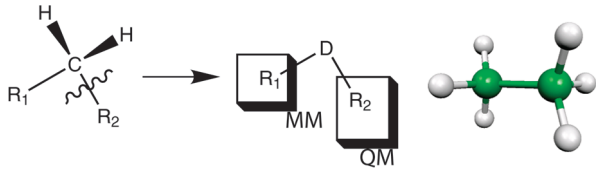


Figure 1. General principle of the repartitioning scheme for a QM/MM calculation in which a chemical bond (here C−R_2) crosses the QM/MM border and is hence cleaved. The link atom that saturates the resulting dangling bond (•R_2) is denoted D. Also shown is the ethane molecule, which serves as the reference molecule for optimizing the parameters of the pseudopotential by which the dummy atom is implemented.

in Figure 1. Similar problems arise when MM atoms are located near a QM region, because the QM and MM descriptions are not genuinely compatible. Thus, a suitable interface has to be used, which can mutually couple the two schemes in a realistic way.

In this work, we address the perturbing effect of a bond cleavage that occurs if a part of a given molecule is treated using quantum mechanics and another part is modeled classically (Figure 1). There are several approaches for tackling such a situation, where most commonly a carbon−carbon bond is cut. In the following, we will refer to the resulting pseudoatom synonymously as the "dummy atom" or capping potential (the "D" atom in Figure 1). There are many QM/MM implementations already available in quantum chemistry software packages; many groups have further developed specific improvements to the QM/MM idea.[15,16,18–25] In particular, there are several approaches to tackle the bond saturation problem arising from a bond cleavage by the QM/MM repartitioning as mentioned above. Among them are the following.

(a) Hydrogen capping: The dummy atom in Figure 1 is represented by a regular hydrogen atom.[26] This relatively straightforward solution has known disadvantages, but it is nevertheless used very often. Obviously, the C−H bond length is shorter than the original C−C bond, and the vibrational frequencies are different. The smaller electronegativity of the hydrogen furthermore changes the electronic structure of the quantum subsystem in the vicinity of the border region considerably. This perturbation can reach over several C−C bonds in the QM subsystem.

(b) Fluorine capping: The saturation of the dangling bond is done via a seven-valent termination atom, for instance, a fluorine, instead of a hydrogen. While this solution, which was originally developed as a pseudobond approach,[23] provides a better bond distance agreement ($d_{C-F} \approx d_{C-C}$), the electronegativity of fluorine is significantly higher than that of carbon. Thus, the electronic subsystem can be perturbed somewhat stronger compared to that with hydrogen capping.

(c) Frozen orbitals: An alternative method relies on precomputed atomic orbitals that are placed at the link atom to ensure an adequate electrostatic interaction and an accurate orthogonality of the terminal chemical bond of the QM subsystem.[27] This frozen-orbital scheme has also been employed for the calculation of NMR shielding constants.[28] A related approach has been developed by the Truhlar group, where auxiliary hybrid orbitals are used to provide an optimal directionality of the termination of the last QM bond.[29,30] While this class of approaches is one of the more accurate ones, it involves a higher coding effort for the incorporation of the frozen orbitals, even though they are excluded from the actual SCF optimization.

(d) Effective fragment potential: Originally designed as a discrete solvation approach to treat chemical reactions in solution,[31] it has been extended to study covalently bound clusters and bulk properties.[32–35] In this method, the total system is divided into a QM region and the environment (the fragment) which interacts with the QM region via a set of one-electron potentials. All important physical interactions between the two fragments (which can be either covalently or noncovalently bonded) are considered explicitly, in particular electrostatic interactions, charge penetration, and polarization effects. Also the effect of exchange repulsion can be incorporated into the scheme. While this effective fragment potential provides a highly accurate description of the original quantum-mechanical interactions, it is not designed to be transferable between different types of fragments. Furthermore, it requires a considerable additional effort for both the design and implementation of the fragment potentials, and it increases the computational effort at runtime compared to conventional QM/MM approaches based on empirical force fields for the MM part. A method that has similar characteristics is known as the effective group potential method,[36,37] but has been used less frequently than the original effective fragment potential approach.

(e) Field-adapted adjustable density matrix assembler (FA-ADMA): A related technique exists in which the target macromolecule is divided into fragments for which conventional quantum chemical calculations are performed.[38–40] Both the fragment and its local environment up to a certain distance are included in these calculations, and the rest of the macromolecule is incorporated via point charges. This approach is hence a regular QM/MM method, with the difference that the QM region is made somewhat larger than really necessary to remove the problems related to the QM/MM boundary region.

(f) Quantum capping potentials: The saturation of dangling bonds with effective potentials has already been attempted by DiLabio et al.[41–45] in an approach that is similar to the one proposed in this work. A conventional pseudopotential is used to truncate the quantum region, using a local part and nonlocal angular-momentum-dependent projectors. These effective capping potentials, however, are not specifically tuned to reproduce the full-QM spectroscopic properties in the QM/MM calculations. Instead, they are built in analogy to the generation of regular atomic pseudopotentials, focusing on the capping atom's orbitals and their energy levels.

In this work, we go one step beyond the QM/MM capping approaches presented above, by using specially designed capping potentials. We present the results of an optimization scheme designed to improve such special potentials within a density functional theory based approach. Specifically, our work is based on analytical effective core potentials (pseudopotentials) of the Goedecker type,[46,47] in line with previous

QM/MM studies.[10,16,48] Our goal is to optimize the pseudopotential parameters in such a way that the change of the electronic density in the quantum part of a QM/MM calculation is minimal with respect to a "full-QM" calculation. In this way, we also ensure that structural parameters and spectroscopic properties in the direct neighborhood of a QM/MM bond cleavage are modeled with a high degree of reliability.

To achieve this aim, we define a penalty functional that quantifies the deviation of the electronic density in a molecular fragment from the corresponding density in the complete molecule, while simultaneously penalizing changes in the equilibrium bond distance and frequency. The penalty functional is minimized iteratively by varying the coefficients of the capping potential placed at the bond cleavage site. This approach is similar to the recently developed heptavalent potential,[49] where we variationally optimized effective atom-centered potentials to describe the methyl group in acetic acid. However, we found that this potential does not always optimally reproduce the spectroscopic parameters of the full molecule. Our capping potentials can be used as link atoms replacing a carbon and involve no further external geometry constraints. They also give reasonable results for potential energy surfaces of the C−C bond. We characterize the perturbative effect of the bond cleavage by means of NMR chemical shifts, which are known to be particularly sensitive to both the intramolecular electronic structure and intermolecular effects such as hydrogen bonding.[50–54] Hence, we can not only gauge the direct perturbing effect of the cleaved bond on the electronic structure of the remaining part of a molecule, but also quantitatively describe how strongly its response properties are tainted by the QM/MM bond cleavage.

2. Methods and Computational Details

2.1. Goal of the Optimization. The purpose of dummy atoms in QM/MM calculations is to enable a saturation of the last covalent bond of the quantum region, i.e., the bond which is cleaved by the QM/MM repartitioning. The central difficulty regarding the quantum region is that the true character of the bond cannot be reproduced easily by a simple terminal atom. Especially spectroscopic parameters react very sensitively to small deviations in the electronic structure around the cleaved bond.

The aim of our optimization scheme is to provide a tool which allows tuning of the properties of the terminal dummy atom in such a way as to make the electronic density in the QM part of the molecule (ρ_D) as similar as possible to the reference electron density (ρ_{QM}), i.e., the density when the entire molecule is treated quantum mechanically. This will eventually lead to an improvement in the spectroscopic properties of the system in the QM/MM description. We further aim at preserving the C−C equilibrium bond length in the dummy calculation to allow an easy coupling of the "first" classical MM atom and to avoid the need for additional geometric constraints.

To this aim, we define a penalty functional which expresses the deviation of these properties from their target values obtained in a full-QM calculation via

$$\mathscr{P} = \int_{\Omega} \mathrm{d}^3 r\, [\rho_{\mathrm{QM}}(\mathbf{r}) - \rho_{\mathscr{D}}(\mathbf{r})]^2 + \sum_{J}^{N_{\mathrm{geom}}} \left\{ w_F \sum_{I}^{N_{\mathrm{ions}}} [F_I^{\mathrm{QM}}(R_J) - F_I^{\mathscr{D}}(R_J)]^2 + w_E[\Delta E^{\mathrm{QM}}(R_J) - \Delta E^{\mathscr{D}}(R_J)]^2 \right\} \quad (1)$$

The integration volume Ω is used to restrict the penalization region to areas in which an improvement is physically meaningful. In our case, this volume corresponds to the union of spheres of 1 Å radius around all QM atoms except the carbon which immediately follows the dummy atom. This definition ensures that the covalent dummy−carbon bond is *not* included in the penalty integration volume, while all other bonds of the first carbon are fully incorporated. w_F and w_E are weighting factors to ensure that an adequate relative importance is given to deviations of the electronic density and the forces and total energy, respectively. Several different molecular geometries (here $N_{\mathrm{geom}} = 3$) are incorporated into the force and energy terms of eq 1 to ensure that not only the equilibrium configuration of the molecule is taken into account. Typically, these conformations will correspond to variations of the bond length of the capping potential.

2.2. Functional Form of the Capping Potential. The capping potentials are represented in the form of analytical effective core potentials of the Goedecker type,[47,49] consisting of a local and a nonlocal part. For a carbon atom, the local potential reads

$$V_{\mathrm{loc}}(\mathbf{r}) = \frac{-Z_{\mathrm{ion}}}{|\mathbf{r}|} \mathrm{erf}[\rho] + \mathrm{e}^{-\rho^2}(C_1 + C_2\rho^2) \quad (2)$$

with the reduced radius $\rho = |\mathbf{r}|/2^{1/2}r_{\mathrm{loc}}$ and the valence charge Z_{ion}, which would be $Z_{\mathrm{ion}} = 4e$ for a regular carbon pseudopotential. The local radius r_{loc} characterizes both the Gaussian smearing of the nuclear charge density resulting in the error function and the decay of the local potential in eq 2. The nonlocal part of the carbon capping potential consists of one s-type and one p-type projector:

$$V_{\mathrm{nl}}(\mathbf{r},\mathbf{r}') = h_{\mathrm{s}} \frac{1}{2\pi^2 r_{\mathrm{s}}^3} \exp\left(-\frac{\mathbf{r}^2 + \mathbf{r}'^2}{2r_{\mathrm{s}}^2}\right) + h_{\mathrm{p}} \frac{32}{225\pi} \frac{\mathbf{r}\mathbf{r}'}{r_{\mathrm{p}}^5} \exp\left(-\frac{\mathbf{r}^2 + \mathbf{r}'^2}{2r_{\mathrm{p}}^2}\right) \sum_{m=0,\pm1} \bar{Y}_1^m(\hat{\mathbf{r}}) Y_1^m(\hat{\mathbf{r}}') \quad (3)$$

with additional characteristic radii r_{s} and r_{p} and the amplitudes h_{s} and h_{p} of one s-type and one p-type projector, respectively. The starting point for the optimization of the capping potential parameters (C_1, C_2, r_{loc}, r_{s}, h_{s}) was the regular carbon pseudopotential with an adjusted valence charge ($Z_{\mathrm{ion}} = 1$).

2.3. Optimization Scheme. Common effective core potentials are often generated by means of a direct inversion of the electronic Schrödinger equation for an isolated atom, with the help of its all-electron orbitals.[55] An alternative approach

consists in iteratively minimizing a penalty functional that expresses the deviations of the pseudo wave function from its all-electron counterpart; this method is commonly used for analytic potentials of the Goedecker type.[46,47]

In analogy to this concept, we optimize our potentials by an iterative Nelder−Mead downhill simplex minimization[56] of the penalty function in eq 1. All seven parameters of the analytic expression in eqs 2 and 3 are varied until the penalty functional becomes stationary. While the derivative of the force and energy terms of the penalty functional with respect to the capping parameters is done via a three-point finite difference, the derivative of the density deviation is done analytically via perturbation theory. On the example of the radius of the s-channel of the potential, this can be achieved according to

$$\frac{\mathrm{d}\mathscr{P}}{\mathrm{d}r_s} = 2\int_\Omega \mathrm{d}^3r\,[\rho_{QM}(\mathbf{r}) - \rho_{\mathscr{D}}(\mathbf{r})]\,\frac{\mathrm{d}\rho_{\mathscr{D}}(\mathbf{r})}{\mathrm{d}r_s} + \ldots \quad (4)$$

in which the term $\mathrm{d}\rho_{\mathscr{D}}(\mathbf{r})/\mathrm{d}r_s$ is computed as the first-order density response of the system with respect to the "perturbation" that is induced by varying the s-channel radius r_s in the capping potential. In this context, $\mathscr{H}^{(1)} = \mathrm{d}V_{nl}/\mathrm{d}r_s$ represents a perturbation Hamiltonian, as would be an external electric or magnetic field in the case of an external perturbation.[57–59]

2.4. Computational Details. Our calculations are done within density functional theory[60–62] using the BLYP[63,64] exchange-correlation functional, as implemented in the CPMD package.[65,66] We use standard norm-conserving pseudopotentials[46,47] and a 70 Ry energy cutoff for the plane-wave expansion of the Kohn−Sham orbitals. To simplify the problem of the bond cleavage and to eliminate the corresponding degrees of freedom, we have not assigned any point charges to the atoms in the classical fragments.

The calculation of magnetic resonance properties (NMR chemical shifts) are done within density functional perturbation theory as implemented in the CPMD package.[57,59,67] Following the experimental convention, we quote chemical shifts relative to computed nuclear shieldings of standard reference systems tetramethylsilane and nitromethane for ^{13}C, ^{1}H, and ^{15}N according to eq 5; all sp^2-hybridized carbons are actually referenced indirectly to TMS via the experimental shift and the computed shieldings of benzene ($\delta^{exptl}_{(C_6H_6)} = 128.4$ ppm[68]) according to eq 6. Chemical shift anisotropies were computed as $\Delta\sigma = \sigma_{33} - {}^1/_2(\sigma_{11} + \sigma_{22})$ using the convention $\sigma_{11} < \sigma_{22} < \sigma_{33}$ for the principal values σ_{ii} of the nuclear shielding tensor.

$$\delta^{calcd}_{(X)} = \frac{1}{3}\,\mathrm{Tr}[\sigma^{calcd}_{(TMS/NMe)} - \sigma^{calcd}_{(X)}] \quad (5)$$

$$\delta^{calcd}_{(X)} = \delta^{exptl}_{(C_6H_6)} + \frac{1}{3}\,\mathrm{Tr}[\sigma^{calcd}_{(C_6H_6)} - \sigma^{calcd}_{(X)}] \quad (6)$$

3. Results and Discussion

3.1. Capping Potentials in the Reference Molecule. The ethane molecule serves as our reference molecule for the optimization of the pseudopotential parameters of the carbon dummy atom. As C−C bonds are the most common bond type within biomolecules, a controllable way of cutting is highly desirable. For the optimization process, one of the methyl groups is replaced by a capping pseudopotential, whose parameters are varied until the penalty functional in eq 1 becomes stationary. We used three different geometries corresponding to a stretching and shortening of the C−$\mathscr{D}$ bond by ±0.24 Å in the penalty functional. Together with relative weights of $w_F = w_E = 1$, this setup turned out to yield a good compromise between geometric and electronic properties. Figure 2 shows the evolution of the parameters r_{loc}, r_s, and r_p during the progress of the optimization.

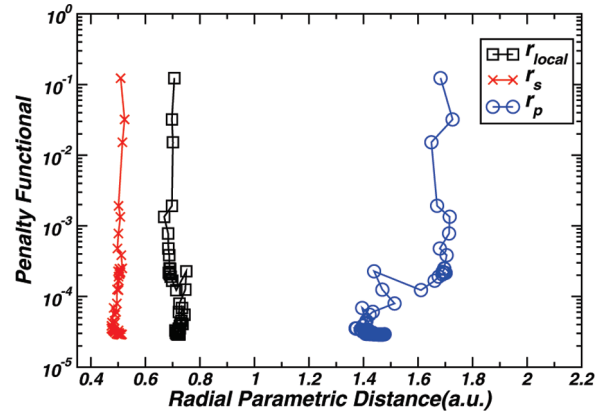


Figure 2. History of the penalty function depending on r_{loc}, r_s, and r_p.

Table 1. Pseudopotenial Parameters of the Regular Carbon Atom and the Optimized Dummy Capping Potential ($\mathscr{D}_C$)[a]

	r_{loc}	C_1	C_2	r_s	h_s	r_p	h_p
regular C	0.3376	–9.1285	1.4251	0.3025	9.6507		
$\mathscr{D}_C$	0.7221	9.9068	–2.5466	0.5120	–3.5081	1.4664	0.2316

[a] Length parameters (r_{loc}, r_s, r_p) are given in units of bohr and energy coefficients (C_1, C_2, h_s, h_p) in hartrees.

The final results for the optimized parameters of the capping potential are shown in Table 1. Note that also it turns out that these values have strongly changed compared with the original carbon pseudopotential from which the optimization was started. Note that the valence charge has been switched from $Z_v = 4$ to $Z_v = 1$. The increase of r_{loc} corresponds to a considerably broader Gaussian smearing of the nuclear charge density, reaching far into the covalent bonding region. The positive C_1 and the negative C_2 coefficients have the effect of further pushing the bonding electron away from the dummy position and attracting it to the middle of the C−$\mathscr{D}$ bond; together, these changes can be seen as a considerably reduced electronegativity of the dummy. The s-channel projector in turn has become attractive, which somewhat compensates the repulsive effect of the local potential (C_1 and C_2). Note that the original carbon pseudopotential had no projector in the p-channel.

3.2. Improvement of Electronic Densities with $\mathscr{D}_{opti}$. In Figure 3, the improvements obtained due to the optimization process for ethane are illustrated in terms of electron density differences. We have compared the density in the full molecule to the density of the dummy-substituted one, using the initial values for the pseudopotential ($\mathscr{D}_{ini}$) and the optimized capping parameters ($\mathscr{D}_{opti}$). We recall that our initial values are the pseudopotential parameters for a regular

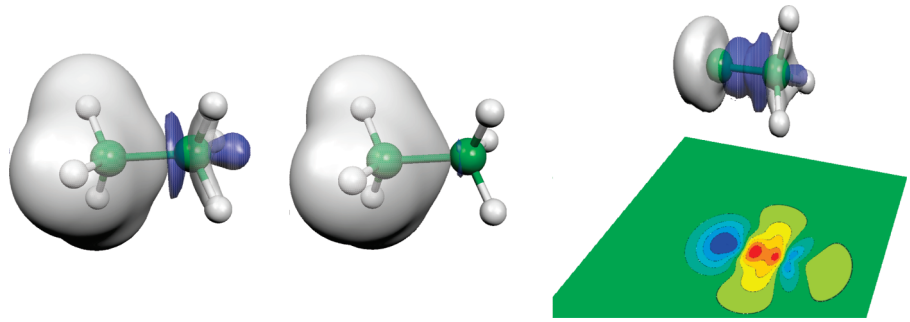


Figure 3. Electron density difference maps for two dummy-substituted ethane molecules. Left: $\rho^{\mathscr{D}_{ini}-CH_3} - \rho^{H_3C-CH_3}$; center: $\rho^{\mathscr{D}_{opti}-CH_3} - \rho^{H_3C-CH_3}$; right: $\rho^{\mathscr{D}_{opti}-CH_3} - \rho^{\mathscr{D}_{ini}-CH_3}$.

Table 2. Calculated ^{1}H and ^{13}C NMR Chemical Shifts δ and Anisotropies $\Delta\sigma$ (ppm) of Ethane before and after the Substitution of the Methyl Group by Dummy Atoms[a]

	(full-QM) R → CH_3	R → $\mathscr{D}_{ini}$	R → $\mathscr{D}_{7v}$	R → $\mathscr{D}_{opti}$	R → H	exptl[69]
δ^H	1.13	–0.72	3.34	0.94	2.21	0.86
$\Delta\sigma^H$	8.6	11.4	13.1	7.9	7.1	
δ^C	10.97	–22.83	–0.31	11.68	28.88	7.00
$\Delta\sigma^C$	17.7	9.4	10.6	23.61	49.0	

[a] In addition to our initial and optimized capping potentials ($\mathscr{D}_{ini}$ and $\mathscr{D}_{opti}$), the heptavalent potential developed by Lilienfeld et al.[49] ($\mathscr{D}_{7v}$) and a simple hydrogen atom were used.

carbon atom (except for the valence charge, which is reduced to 1). Finally, a direct comparison of $\mathscr{D}-CH_3$ between the initial and optimized dummy link atom $\mathscr{D}$ is shown in Figure 3, along with its projection in two dimensions.

When the unoptimized dummy $\mathscr{D}_i$ is used, the deviations of the electronic density with respect to the corresponding full-QM calculation reach somewhat beyond the cleavage bond ($\mathscr{D}-C$), with regions of both increased and decreased electron density (blue and white clouds in Figure 3). The optimized dummy yields somewhat lower density differences with respect to the unperturbed molecule beyond the first regular carbon atom. When comparing the density differences between the initial and optimized dummy atoms directly (rightmost plot in Figure 3), the strongest effect is located at the dummy itself. Nevertheless, also at the methyl protons, the density redistribution is still considerable.

3.3. NMR Chemical Shifts with the Optimized Capping Potential. We have benchmarked the accuracy of the optimized dummy atoms by calculating NMR chemical shifts, which represent the electronic response to an external magnetic field. These NMR parameters offer a unique reduction of the complex electronic structure in the vicinity of a nucleus into a single number, and they are highly sensitive to small changes in the electronic orbitals. In this way, they offer a local orbital-based probe, complementary to the penalty functional itself that is based only on the total density and geometric quantities. The isotropic NMR chemical shifts of the dummy-substituted molecule are compared to those of the reference molecule in Table 2. Both the optimized ($\mathscr{D}_{opti}$) and the unoptimized ($\mathscr{D}_{ini}$) monovalent dummy potentials are used, as well as the seven-valent one ($\mathscr{D}_{7v}$) by Lilienfeld et al.,[49] which was developed to reproduce the electronic density in acetic acid. For the sake of completeness, we have also added the results for a simple hydrogen capping. To exclude the effects of conformational changes on the NMR chemical shifts, we have always used the optimized geometries of the full molecule.

Table 3. Optimized Bond Lengths (Å) and Vibrational Frequencies of the C−$\mathscr{D}$ Bond for the Ethane Reference Molecule before and after Substitution of the Methyl Group

	(full-QM) R → CH_3	R → $\mathscr{D}_{ini}$	R → $\mathscr{D}_{opti}$	R → $\mathscr{D}_{7v}$	R → H
d(R−C) (Å)	1.54	1.68	1.54	2.04	1.10
d(C−H) (Å)	1.10	1.10	1.10	1.09	1.10
θ_{H-C-R} (deg)	111.3	110.3	112.8	106.8	109.4
θ_{H-C-H} (deg)	107.6	108.6	105.9	112.0	109.4
$\tilde{\nu}_{C-\mathscr{D}}$ (cm^{-1})	909	483	809	682	1103

It turns out (see Table 2) that the chemical shifts from our optimized capping atom are generally in better agreement with the all-QM calculation than for the initial ($\mathscr{D}_{ini}$) and seven-valent ($\mathscr{D}_{7v}$) capping potentials. The initial potential results in significantly lower chemical shifts, while the heptavalent dummy overestimates the ^{1}H shifts and underestimates the ^{13}C shifts. Hydrogen capping in turn always results in too positive chemical shifts. Only the optimized monovalent substitution yields values close to the all-quantum calculation for both nuclei. The deviations between the reference and $\mathscr{D}_{opti}$ are below 0.3 ppm for protons and 1 ppm for carbon, which are errors that can certainly be tolerated for nuclei that are only one or two bonds away from the capping atom. This is not the case for all other capping variants, which exhibit deviations in the NMR chemical shifts of more than 1 ppm (^{1}H) and 20−30 ppm (^{13}C). In conclusion, the NMR resonances of our dummy-substituted ethane show indeed a very good agreement with the corresponding full-quantum calculations, even for the atom directly connected to the bond cleavage.

3.4. Energetic and Geometric Properties of the $\mathscr{D}$−C Bonds. We have optimized the geometry of the reference ethane with the methyl group substituted by the optimized dummy atom. The results for selected distances and angles are shown in Table 3, and Figure 4 shows the potential energy profile of ethane as a function of the $\mathscr{D}$−C

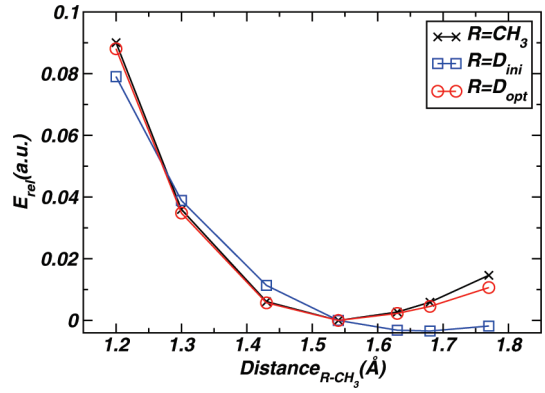


Figure 4. Potential energy curve for rigid stretching/compression of the C−C bond in ethane and $\mathscr{D}$−CH_3 with the original and the optimized pseudopotentials.

bond length. The equilibrium bond length of $\mathscr{D}_{ini}$−CH_3 is somewhat longer than the full QM value. The optimized capping potential $\mathscr{D}_{opti}$, in turn, improved this distance considerably. Also compared to the heptavalent dummy ($\mathscr{D}_{7v}$), we find a clear improvement of the equilibrium bond distances, which is most likely due to the incorporation of the atomic forces into our new penalty functional, eq 1.

Last but now least, the harmonic stretch frequency of the $\mathscr{D}_{opti}$−C bond is almost perfectly reproduced, while the $\mathscr{D}_{ini}$−C and $\mathscr{D}_{7v}$−C variants show deviations. Not surprisingly, the harmonic frequency of the hydrogen capping cannot accurately reproduce the C−C frequency either (although the carbon mass was used instead of the hydrogen one in the calculation of the dynamical matrix).

The full potential energy curve for the $\mathscr{D}$−C bond is shown in Figure 4. The agreement of the optimized dummy potential (red) with the full-QM dissociation curve (black) is remarkable, while the initial capping potential shows an underestimated dissociation energy and a somewhat extended equilibrium bond length. For comparison, we also optimized the capping potential using somewhat lower weights for the geometric penalty contributions (w_F and w_E, data not shown). This calculation resulted in a capping potential that was numerically more similar to the original one (the regular C potential; see Table 1), at the expense of a considerably worse C−$\mathscr{D}$ bond distance and vibrational frequency.

3.5. Histidine. The good agreement obtained in the previous section might have been fortuitous, as the dummy potentials were optimized for the very specific molecule that was subsequently benchmarked there. Thus, we have checked the transferability of our dummy potentials by applying them to a different molecule, namely, histidine. Two selected bond lengths are listed in Table 4, comparing the initial and final capping potentials to the full-QM results. In analogy to the situation encountered for the ethane molecule, the initial $\mathscr{D}_{ini}$ capping yields a stretched $\mathscr{D}$−C_2 bond length, while C_2−C_3 is slightly shortened. Both deficiencies are considerably improved upon by $\mathscr{D}_{opti}$.

Table 5 shows the NMR chemical shifts of the full histidine molecule and its imidazole fragment within a QM/MM description, always using the optimized geometry of the full histidine molecule. As expected, the strongest deviations are observed for carbon C_2 directly involved in the bond cleavage. Here, the capping optimization scheme results in the reduction of the error by almost an order of magnitude compared to the unoptimized capping atom. Similarly, a considerable improvement is obtained for the next carbons C_3 and C_6, as well as nitrogen N_7. In all these cases, the optimized monovalent potential also performs better than the heptavalent dummy atom. For C_9 and N_8, which are further away from the bond cleavage, the situation

Table 4. Geometric Data as Well as Computed and Experimental Proton Affinities $\Delta E = E_{DFT}(X) - E_{DFT}(X{-}H^+)$ for Histidine and Lysine as Well as their Dummy-Substituted Fragments[a]

	full-QM	R → $\mathscr{D}_{ini}$	R → $\mathscr{D}_{opti}$	R → H
		Histidine		
$d_{\mathscr{D}-C_2}$ (Å)	1.57	1.69	1.55	1.1
$d_{C_2-C_3}$ (Å)	1.51	1.49	1.51	1.50
ΔE (kcal/mol)	238.1	245.8	239.4	235.8
		Lysine		
$d_{\mathscr{D}-C_1}$ (Å)	1.56	1.69	1.55	1.1
$d_{C_1-C_2}$ (Å)	1.55	1.55	1.55	1.55
ΔE (kcal/mol)	225.6	232.8	228.3	76.96

[a] For the atom numbering, see Figure 5. R represents the classical part of the molecule, i.e., the amino and carboxylic acid groups.

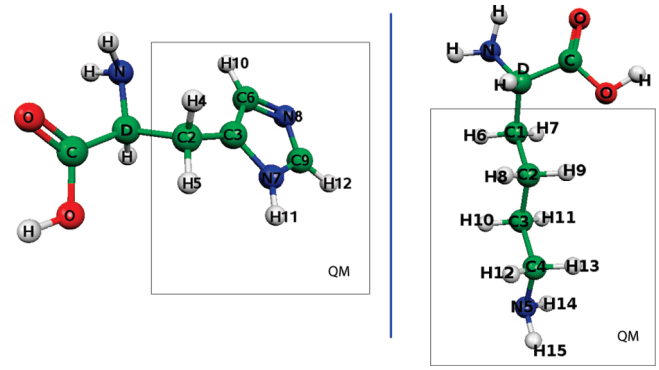


Figure 5. Atom numbering and bond cutting scheme for the histidine (left) and lysine (right) molecules. The molecular fragments outside the "QM" boxes are substituted by means of the capping potential (the dummy atom).

Table 5. Calculated ^1H, ^{13}C, and ^{15}N NMR Chemical Shifts (ppm) of Histidine and Its Dummy-Substituted Fragment[a]

	full-QM	R → $\mathscr{D}_{opti}$	R → $\mathscr{D}_{ini}$	R → $\mathscr{D}_{7v}$	R → H	exptl[69]
C_2	32.46	27.03	−2.72	11.11	43.03	30.78
$\Delta\sigma(C_2)$	25.72	27.05	41.46			
C_3	136.90	140.90	142.05	141.54	137.35	134.67
$\Delta\sigma(C_3)$	80.40	84.40	84.04			
C_6	137.37	133.65	132.52	135.03	138.45	119.55
C_9	138.05	136.37	135.96	137.37	137.85	138.97
N_7	−202.97	−202.74	−200.70	−204.79	−202.73	-
N_8	−138.91	−139.26	−139.60	−139.60	−137.12	-
H_4	3.40	3.66	2.62	5.99	5.01	3.12
$\Delta\sigma(H_4)$	6.39	7.84	10.09			
H_5	3.56	3.14	2.03	5.26	4.46	3.23
$\Delta\sigma(H_5)$	6.53	6.73	8.63			
H_{10}	7.60	6.86	6.56	7.28	7.06	7.06
H_{11}	9.93	9.82	9.66	9.93	9.84	
H_{12}	7.33	7.17	7.09	7.20	7.25	7.80

[a] For the atom numbering, see Figure 5. Data for both the initial and optimized dummy potentials ($\mathscr{D}_{ini}$ and $\mathscr{D}_{opti}$) as well as the heptavalent one ($\mathscr{D}_{7v}$)[49] are shown.

Table 6. Calculated ^{1}H, ^{13}C, and ^{15}N NMR Chemical Shifts (ppm) of Lysine and Its Dummy-Substituted Fragment[a]

	full-QM	R → $\mathscr{D}_{ini}$	R → $\mathscr{D}_{opti}$	R → H	exptl[69]
C_1	37.54	7.33	37.27	53.11	32.60
$\Delta\sigma(C_1)$	35.69	40.08	38.11		
C_2	28.81	30.88	30.79	28.66	24.13
C_3	35.27	43.81	36.98	35.15	29.11
C_4	48.10	48.27	47.81	47.58	41.75
N_5	–293.21	–292.82	–292.95	–293.25	
$H_{6,7}$	2.25	1.13	2.24	3.56	1.9
$\Delta\sigma(H_{6,7})$	4.17	7.88	6.04		
$H_{8,9}$	2.25	2.13	2.24	2.36	1.5
$H_{10,11}$	2.10	1.99	1.91	1.93	1.7
$H_{12,13}$	3.72	3.59	3.65	3.69	3.0
H_{14}	1.95	1.88	1.90	1.93	
H_{15}	2.78	2.70	2.74	2.77	

[a] For the atom numbering, see Figure 5. R represents the remaining fragments in the classical regions.

is less drastic, and all three choices yield similar—yet small—discrepancies with respect to the full-QM calculation.

The hydrogens exhibit smaller absolute deviations, which is because their NMR chemical shift spectrum spans a range that is about an order of magnitude smaller than that of C and N nuclei. Nevertheless, the shifts of the hydrogens adjacent to the bond cut (H_4 and H_5) are considerably better. The problem for H_4 might be due to the effect of the lone electron pairs of the nearby nitrogen, which are missing entirely when the left fragment is replaced by the capping potential. We think that the same phenomenon applies to H_{10}, which is in somewhat better agreement after the optimization. The deviations of H_{11} and H_{12} are quite small and do not obviously correlate to the choice of the capping potential. In most cases, the monovalent capping potential outperforms the heptavalent one.

In addition to the NMR chemical shifts, we have used the optimized capping potential to compute the proton affinities, with N_8 as the protonation site. The energy differences between the neutral and charged histidine molecules (also given in Table 4) from the $\mathscr{D}_{opti}$ capping model are quite accurate when compared to the full-QM results. The error from the full-QM proton affinity is reduced from $\Delta E = 7.7$ kcal/mol to $\Delta E = 1.3$ kcal/mol, which corresponds to less than 1% of the total affinity. In conclusion, the capping potential which was optimized for the ethane molecules yields a very good overall accuracy when transferred to histidine.

3.6. Lysine. Our second QM/MM application of the capping potential is the amino acid lysine, which was split into QM/MM fragments as illustrated in Figure 5. In our setup, the amino group was replaced by our initial and optimized dummy potentials; note that also here the $\mathscr{D}_{opti}$ parameters were taken from the ethane-based optimization without any further change. The results of the geometry optimization using our initial and optimized capping potentials $\mathscr{D}_{ini}$ and $\mathscr{D}_{opti}$ are shown in Table 4. While the C_1-C_2 bond is not affected by the bond cleavage, the $\mathscr{D}-C_1$ bond length is notably different for the unoptimized capping potential. A similar improvement is observed for the proton affinity of lysine.

The NMR chemical shifts obtained via our full-QM and QM/MM calculations are summarized in Table 6, again using the optimized geometry of the full lysine. It turns out that the results for the $\mathscr{D}_{opti}$ capping potential are in better agreement with the full-QM values than that those from the hydrogen capping and the unoptimized dummy $\mathscr{D}_{ini}$. The deviation of the C_1 chemical shift is decreased from 30 to <1 ppm, and for H_6 and H_7, we find improvements from $\Delta\delta = 1.1$ ppm to $\Delta\delta < 0.1$ ppm. This suggests that the nonoptimized dummy potential leads to a distortion of the electronic density that significantly affects the atoms near the "broken" QM/MM bond. Inspecting the shifts of C_2, C_3, and C_4 under the $\mathscr{D}_{ini}$ capping, we find that the influence of the bond cleavage is not negligible even for the atoms which are located several bonds away from the QM/MM border. Nevertheless, the new $\mathscr{D}_{opti}$ agrees quite well with the full-QM calculations for all atoms.

4. Conclusion

We have presented improved monovalent capping potentials for hybrid QM/MM calculations within density functional theory. The parameters of analytic effective pseudopotentials are optimized such as to reproduce the electronic density, proton affinities, atomic forces, and geometries as closely as possible with respect to the corresponding full-QM quantities. Particular focus is put on the reliability of NMR chemical shifts as highly sensitive probes of the ground-state and response properties of the electronic orbitals. The resulting analytic capping potentials are shown to have a high transferability for different molecules. An important advantage resulting from the improved electronic structure of the optimized capping potentials is that our $\mathscr{D}_{opti}$ can help to reduce the QM box size significantly, since the perturbation of the QM/MM bond cleavage is essentially undetectable in the QM region beyond one single chemical bond.

Acknowledgment. S.K. thanks the Ubon Ratjatanee University. D.S. acknowledges support from the Deutsche Forschungsgemeinschaft (DFG) under Grants SE-1008/5 and SE-1008/6. We are grateful to Jochen Schmidt for useful discussions and for his precious help.

References

(1) Lehn, J.-M. *Rep. Prog. Phys.* **2004**, *67*, 249–265.

(2) Tolstoy, P. M.; Schah-Mohammedi, P.; Smirnov, S. N.; Golubev, N. S.; Denisov, G. S.; Limbach, H. H. *J. Am. Chem. Soc.* **2004**, *126*, 5621–5634.

(3) Meng, S.; Xu, L. F.; Wang, E. G.; Gao, S. *Phys. Rev. Lett.* **2002**, *89*, 176104.

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

(5) Rapp, A.; Schnell, I.; Sebastiani, D.; Brown, S. P.; Percec, V.; Spiess, H. W. *J. Am. Chem. Soc.* **2003**, *125*, 13284–13297.

(6) Goward, G.; Sebastiani, D.; Schnell, I.; Spiess, H. W. *J. Am. Chem. Soc.* **2003**, *125*, 5792–5800.

(7) Lee, Y.; Murakhtina, T.; Sebastiani, D.; Spiess, H. *J. Am. Chem. Soc.* **2007**, *129*, 12406–12407.

(8) Gervais, C.; Dupree, R.; Pike, K. J.; Bonhomme, C.; Profeta, M.; Pickard, C. J.; Mauri, F. *J. Phys. Chem. A* **2005**, *109*, 6960–6969.

(9) Murakhtina, T.; Delle Site, L.; Sebastiani, D. *ChemPhysChem* **2006**, *7*, 1215–1219.

(10) Rohrig, U.; Guidoni, L.; Laio, A.; Frank, I.; Rothlisberger, U. *J. Am. Chem. Soc.* **2004**, *126*, 15328–15329.

(11) Sebastiani, D.; Parrinello, M. *ChemPhysChem* **2002**, *3*, 675.

(12) Murakhtina, T.; Heuft, J.; Meijer, J.-E.; Sebastiani, D. *ChemPhysChem* **2006**, *7*, 2578–2584.

(13) Bühl, M.; Grigoleit, S.; Kabrede, H.; Mauschick, F. T. *Chem.—Eur. J.* **2006**, *12*, 477–488.

(14) Hansen, M. R.; Sekharan, S.; Graf, R.; Sebastiani, D. *J. Am. Chem. Soc.* **2009**, *131*, 5251–5256.

(15) Deng, R. Z.; Martyna, G. J.; Klein, M. L. *Phys. Rev. Lett.* **1993**, *71*, 267.

(16) Laio, A.; VandeVondele, J.; Roethlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(17) Komin, S.; Gossens, C.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. *J. Phys. Chem. B* **2007**, *111*, 5225–5232.

(18) Wei, D.; Salahub, D. *Chem. Phys. Lett.* **1994**, *224*, 291.

(19) Stanton, R. V.; Little, L. R.; Merz, K. M. *J. Phys. Chem.* **1996**, *99*, 11266.

(20) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1999**, *21*, 10452.

(21) Lyne, P.; Hodosceck, M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 3462.

(22) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

(23) Zhang, Y.; Lee, T.-S.; Yang, W. *J. Phys. Chem.* **1999**, *110*, 46–54.

(24) Brancato, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 144501.

(25) Cui, Q. *J. Chem. Phys.* **2002**, *117*, 4720–4728.

(26) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.

(27) Assfeld, X.; Rivail, J.-L. *Chem. Phys. Lett.* **1996**, *263*, 100–106.

(28) Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2006**, *125*, 194104.

(29) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 632–650.

(30) Jung, J.; Choi, C. H.; Sugita, Y.; Ten-no, S. *J. Chem. Phys.* **2007**, *127*, 204102.

(31) Ohta, K.; Yoshioka, Y.; Morokuma, K.; Kitaura, K. *Chem. Phys. Lett.* **1983**, *101*, 12–17.

(32) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D. *J. Chem. Phys.* **1996**, *105*, 1968.

(33) Adamovic, I.; Freitag, M. A.; Gordon, M. S. *J. Chem. Phys.* **2003**, *118*, 6725–6732.

(34) Netzloff, H. M.; Gordon, M. S. *J. Chem. Phys.* **2004**, *121*, 2711–2714.

(35) Adamovic, I.; Gordon, M. S. *J. Phys. Chem. A* **2006**, *110*, 10267–10273.

(36) Poteau, R.; Ortega, I.; Alary, F.; Solis, A. R.; Barthelat, J.-C.; Daudey, J.-P. *J. Phys. Chem. A* **2001**, *105*, 198–205.

(37) Poteau, R.; Alary, F.; Makarim, H. A. E.; Heully, J.-L.; Barthelat, J.-C.; Daudey, J.-P. *J. Phys. Chem. A* **2001**, *105*, 206–214.

(38) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980–1986.

(39) Exner, T. E.; Mezey, P. G. *Phys. Chem. Chem. Phys.* **2005**, *24*, 4061–4069.

(40) Eckard, S.; Exner, T. E. *Z. Phys. Chem.* **2006**, *220*, 927–944.

(41) Jardilliera, N.; Goursot, A. *Chem. Phys. Lett.* **2008**, *454*, 65–69.

(42) Mallik, A.; Taylor, D. E.; Runge, K.; Dufty, J. W. *Int. J. Quantum Chem.* **2004**, *100*, 1019–1025.

(43) DiLabio, G. A.; Wolkow, R. A.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 044708.

(44) Slavicek, P.; Martinez, T. J. *J. Chem. Phys.* **2006**, *124*, 084107.

(45) DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. *J. Chem. Phys.* **2002**, *116*, 9578–9584.

(46) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703.

(47) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641.

(48) Rohrig, U. F.; Sebastiani, D. *J. Phys. Chem. B* **2008**, *112*, 1267–1274.

(49) von Lilienfeld-Toal, A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 014113.

(50) Brown, S. P.; Spiess, H. W. *Chem. Rev.* **2001**, *101*–4125.

(51) Spiess, H. W. *Macromol. Chem. Phys.* **2003**, *204*, 340–346.

(52) Schulz-Dobrick, M.; Metzroth, T.; Spiess, H. W.; Gauss, J.; Schnell, I. *ChemPhysChem* **2005**, *6*, 315–327.

(53) Ochsenfeld, C.; Brown, S. P.; Schnell, I.; Gauss, J.; Spiess, H. W. *J. Am. Chem. Soc.* **2001**, *123*, 2597–2606.

(54) Bühl, M.; Kabrede, H.; Diss, R.; Wipff, G. *J. Am. Chem. Soc.* **2006**, *128*, 6357–6368.

(55) Pickett, W. E. *Comput. Phys. Rep.* **1989**, *9*, 115–198.

(56) Press, W. H.; Teukoldky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 1992.

(57) Putrino, A.; Sebastiani, D.; Parrinello, M. *J. Chem. Phys.* **2000**, *113*, 7102–7109.

(58) Putrino, A.; Parrinello, M. *Phys. Rev. Lett.* **2002**, *88*, 176401.

(59) Sebastiani, D.; Parrinello, M. *J. Phys. Chem. A* **2001**, *105*, 1951.

(60) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

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

(62) Jones, R. O.; Gunnarsson, O. *Rev. Mod. Phys.* **1989**, *61*, 689–746.

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

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

(65) Hutter, J.; Marx, D.; Focher, P.; Tuckerman, M.; Andreoni, W.; Curioni, A.; Fois, E.; Rothlisberger, U.; Giannozzi, P.; Deutsch, T.; Alavi, A.; Sebastiani, D.; Laio, A.; VandeVondele, J.; Seitsonen, A.; Billeter, S.; Parrinello, M. *Computer Code CPMD*, version 3.12; copyright IBM Corp. and MPI-FKF Stuttgart 1990−2007; http://www.cpmd.org.

(66) Hutter, J.; Curioni, A. *ChemPhysChem* **2005**, *6*, 1788–1793.

(67) Sebastiani, D.; Goward, G.; Schnell, I.; Spiess, H. W. *J. Mol. Struct.: THEOCHEM* **2003**, *625*, 283–288.

(68) Gottlieb, H. E.; Kotlyar, V.; Nudelman, A. *J. Org. Chem.* **1997**, *62*, 7512–7515.

(69) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H.; Markley, J. L. *Nucleic Acids Res.* **2007**, *36*, D402−D408.

CT800525U

JCTC Journal of Chemical Theory and Computation

Non-self-consistent Density-Functional Theory Exchange-Correlation Forces for GGA Functionals

Antonio S. Torralba,*,†,¶ David R. Bowler,† Tsuyoshi Miyazaki,‡ and Michael J. Gillan†

London Centre for Nanotechnology, UCL, 17-19 Gordon St, London WC1H 0AH, U.K., and Materials Simulation Laboratory and Department of Physics and Astronomy, UCL, Gower St, London WC1E 6BT, and National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0045, JAPAN

Received December 6, 2008

Abstract: When using density functional theory (DFT), generalized gradient approximation (GGA) functionals are often necessary for accurate modeling of important properties of biomolecules, including hydrogen-bond strengths and relative energies of conformers. We consider the calculations of forces using non-self-consistent (NSC) methods based on the Harris−Foulkes expression for energy. We derive an expression for the GGA NSC force on atoms, valid for a hierarchy of methods based on local orbitals, and discuss its implementation in the linear scaling DFT code Conquest, using a standard (White−Bird) approach. We investigate the use of NSC structural relaxations before full self-consistent relaxations as a method for improving convergence. Example calculations for glycine and small alanine peptides suggest that NSC pre-relaxations of the structure are indeed useful to save computer effort and time.

1. Introduction

Density functional theory (DFT) has become a standard technique in materials simulations over the last twenty years and is making an increasing impact in theoretical studies of biochemical systems, especially those involving radicals[1] and catalysis by metals.[2] While plane waves are often used as a basis in condensed phases, local orbital basis functions make it natural to work with a hierarchy of methods from non-self-consistent (NSC) DFT calculations based on the Harris−Foulkes (HF) expression for total energy[3,4] to full DFT, as well as providing an ideal framework for linear scaling implementations. This hierarchy gives us the power to perform many exploratory calculations rather quickly at low precision, and then increase the precision in a well-controlled way. The purpose of this paper is to extend our previous treatment of forces[5] for local orbital methods to non-self-consistent GGA calculations, to discuss its implementation in the Conquest linear scaling DFT code[6,7] and to explore one aspect of the hierarchical approach: the efficacy of an initial NSC relaxation in speeding up SC relaxations.

For biological systems requiring ab initio accuracy, typically QM/MM methods are used,[8] though with recent developments in linear scaling DFT,[9] whole molecules or large portions of molecules can be addressed.[10−12] For small QM regions, hybrid functionals can be used, but this is still not practical for large regions despite progress with local Hartree−Fock methods[13] and screened hybrid functionals,[14] so GGA is still important for large biochemical applications[15] and is necessary for properties such as hydrogen-bond strengths and relative energies of conformers.[16] The localization of linear scaling methods also makes them natural candidates for embedding techniques, such as QM/MM, and for embedding full DFT calculations into less precise calculations such as NSC techniques.[17]

We recall the hierarchy of electronic structure methods[5] of different accuracy for clarity. A first possibility is to perform fully self-consistent DFT calculations with a fixed

* To whom correspondence should be addressed. E-mail: TORRALBA.Antonio@nims.go.jp.

† London Centre for Nanotechnology and Materials Simulation Laboratory and Department of Physics and Astronomy.

‡ National Institute for Materials Science.

¶ Present address: National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0045, Japan.

10.1021/ct8005425 CCC: $40.75
Published on Web 05/04/2009

basis of pseudoatomic local orbitals (PAOs).[6,18−21] Even more accurate is to allow the local orbitals to vary by expanding them in a sufficiently complete basis, such as B-splines[22] or periodic sinc functions.[23] We call this full DFT. For lower precision calculations, on the other hand, we can use the non-self-consistent (NSC) Harris−Foulkes functional with a fixed and often limited basis of PAOs and fix the electronic density (often to a superposition of atomic densities).[24] Methods using a limited, fixed PAO basis bear similarities with standard tight binding and, for this reason, are sometimes called ab initio tight binding (AITB). When NSC methods are used, the forces can be written[5] so that there is a clear contribution to the total force from the difference between input and output charge densities; here, we derive a form for this force in which the exchange-correlation functional depends on the gradient of the charge density, as well as the charge itself (as is the case for GGA functionals). We note that the calculation of a correction to the forces on ions caused by incomplete convergence of self-consistency cycle is important for consistency between energy and forces, and has been considered before,[25,26] though not in the context of local orbital or linear scaling methods (which present particular challenges, requiring locality and expressions in terms of density matrices rather than KS orbitals). A scheme to accelerate LDA molecular dynamics using NSC LDA forces and local orbitals has also been proposed.[27]

The paper is laid out as follows: in the next section, we summarize the local orbital approach to DFT, recalling in particular the linear scaling implementation and NSC forces; we then derive the expression for NSC forces with reference to a uniform grid (though the result can be easily extended to other implementations); we then present tests comparing full SC structural relaxation and NSC, followed by SC structural relaxation, and conclude.

2. Local Orbitals and Non-Self-Consistent Forces

As explained in detail elsewhere,[5] we perform DFT calculations using a basis set of local orbitals (known as support functions in Conquest), $\phi_{i\alpha}(\mathbf{r})$, where i denotes an atom and α is a local orbital on the atom. We assume throughout the use of pseudopotentials, though this is not restrictive. The total energy is then written

$$E_{\mathrm{Tot}} = E_{\mathrm{kin}} + E_{\mathrm{ps}} + E_{\mathrm{Har}} + E_{\mathrm{XC}} + E_{\mathrm{c}} \quad (1)$$

where the kinetic (E_{kin}) and pseudopotential (E_{ps}) energies are found as usual, E_{c} is the core−core Coulomb interaction, and the Hartree (E_{Har}) and exchange-correlation (E_{XC}) energies depend on the charge density

$$n(\mathbf{r}) = 2\sum_n f_n |\psi_n(\mathbf{r})|^2 \quad (2)$$

where $\psi_n(\mathbf{r})$ is a Kohn−Sham eigenstate and f_n are orbital occupancies.

The Kohn−Sham eigenstates are written in terms of the local orbitals

$$\psi_n(\mathbf{r}) = \sum_{i\alpha} u^n_{i\alpha}\phi_{i\alpha}(\mathbf{r}) \quad (3)$$

We note that the local orbitals are strictly localized within some cutoff radius, and are in general nonorthogonal ($\langle\phi_{i\alpha}|\phi_{j\beta}\rangle = S_{ij\alpha\beta}$). For local orbitals $\phi_{i\alpha}(\mathbf{r})$, whose form is fixed, the total energy of the system is only a function of the expansion coefficients $u^n_{i\alpha}$ and the ground state is found by minimizing the energy with respect to the coefficients, while imposing orthonormality on the KS eigenstates and self-consistency. This gives self-consistent DFT; full DFT can be achieved by writing the local orbitals in terms of a sufficiently complete basis. If, instead, the charge density is fixed and composed of a superposition of pseudoatomic densities

$$n^{\mathrm{in}}(\mathbf{r}) = \sum_i \eta_i(|\mathbf{r} - \mathbf{R}_i|) \quad (4)$$

with $\eta_i(|\mathbf{r} - \mathbf{R}_i|)$ the spherically symmetric density for atom i which has position $\mathbf{R}_i$, then we have non-self-consistent DFT. We use the Harris−Foulkes energy[3,4] and write

$$E_{\mathrm{Tot}} = E_{\mathrm{BS}} + \Delta E_{\mathrm{Har}} + \Delta E_{\mathrm{XC}} + E_{\mathrm{c}} \quad (5)$$

Here the Hartree and exchange-correlation double counting terms (ΔE_{Har} and ΔE_{XC}) are calculated with $n^{\mathrm{in}}(\mathbf{r})$. In standard DFT, the band-structure energy is given by a sum over KS eigenenergies, $E_{\mathrm{BS}} = \Sigma_i f_i\varepsilon_i$, where f_i is the occupation. The band-structure energy can also be written in terms of a density matrix, ρ, as $E_{\mathrm{BS}} = 2\mathrm{Tr}[\rho H]$. In linear scaling DFT codes,[6,7,19,20,23] the energy is variationally minimized with respect to the density matrix, with an explicit localization constraint imposed (a good approximation for systems with a gap, where there the density matrix decays exponentially with distance).[28] The Harris−Foulkes energy agrees exactly with the standard KS expression (eq 1) at self-consistency, and away from self-consistency deviates from that energy by an amount second order in the deviation of $n^{\mathrm{in}}(\mathbf{r})$ from the self-consistent ground state $n(\mathbf{r})$.

The forces can be conveniently written in terms of the density matrix, whether this is found by exact diagonalization or linear scaling techniques, and the form of the equations is identical, as explained in detail elsewhere.[5] The final form is

$$\mathbf{F}_i = \mathbf{F}^{\mathrm{ps}}_i + \mathbf{F}^{\mathrm{P}}_i + \mathbf{F}^{\mathrm{NSC}}_i + \mathbf{F}^{\mathrm{c}}_i \quad (6)$$

where there are Hellmann−Feynman contributions from the motion of the pseudopotentials (denoted ps), Pulay contributions from the motion of basis functions with the atoms (denoted P; we note that some linear scaling codes use basis functions which do not move with atoms, and do not have these forces), contributions from non-self-consistency (denoted NSC) and finally from ionic core interaction (denoted c).

The only difference between the SC and NSC forms is the NSC force term itself, which is written

$$\mathbf{F}^{\mathrm{NSC}}_i = -\int \mathrm{d}\mathbf{r}[\delta V_{\mathrm{Har}}(\mathbf{r})\nabla_i n^{\mathrm{in}}(\mathbf{r}) + \delta n(\mathbf{r})\nabla_i V_{\mathrm{XC}}(\mathbf{r})] \quad (7)$$

where $\delta n(\mathbf{r}) = n^{\mathrm{out}}(\mathbf{r}) - n^{\mathrm{in}}(\mathbf{r})$ is the difference between the input density and the density built from the KS orbitals (eq 2), $\delta V_{\mathrm{Har}}(\mathbf{r})$ is the Hartree potential because $\delta n(\mathbf{r})$ and $\nabla_i V_{\mathrm{XC}}(\mathbf{r})$ is the gradient of the exchange-correlation potential

for the input density with respect to the position of atom i. The derivation of $\mathbf{F}_i^{\mathrm{NSC}}$ previously given[5] assumed an LDA functional and needs to be reworked for GGA.

GGA functionals take the form

$$E_{\mathrm{xc}}[n(\mathbf{r})] = \int \mathrm{d}\mathbf{r} f_{\mathrm{xc}}(n(\mathbf{r}), \nabla n(\mathbf{r})) \tag{8}$$

Note that this is commonly written in a different but equivalent form, using enhancement factors F_{xc}, so that $f_{\mathrm{xc}} = n\varepsilon_{\mathbf{x}}^{\mathrm{unif}}(n)F_{\mathrm{XC}}(n,\nabla n)$, where $\varepsilon_{\mathbf{x}}^{\mathrm{unif}}$ is the uniform electron gas exchange energy density. The GGA kernel f_{xc} depends explicitly on both the electronic density, $n(\mathbf{r})$, and its gradient, $\nabla n(\mathbf{r})$ (which will be denoted $\mathbf{g}(\mathbf{r})$ for brevity). In almost all GGA functionals the dependence is only on the magnitude of the gradient, $|\nabla n(\mathbf{r})|$; we preserve the full gradient both for completeness and because the algebra is clearer; we use the magnitude of the gradient at the end of the derivation. The gradient of the potential with respect to the position of the atom i needed to compute (eq 7) becomes

$$\nabla_i V_{\mathrm{xc}}(\mathbf{r}) = \mu'_{\mathrm{xc}}[n(\mathbf{r}), \mathbf{g}(\mathbf{r})]\nabla_i n(\mathbf{r}) \tag{9}$$

with $\mu'_{\mathrm{xc}}(n) = \mathrm{d}V_{\mathrm{XC}}/\mathrm{d}n(\mathbf{r})$. The exchange-correlation potential is written[29]

$$V_{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}}{\delta n(\mathbf{r})} = \frac{\partial f_{\mathrm{xc}}}{\partial n(\mathbf{r})} - \nabla\cdot\frac{\partial f_{\mathrm{xc}}}{\partial \mathbf{g}(\mathbf{r})} \tag{10}$$

3. GGA NSC Forces

To derive the GGA NSC force on atoms, we must account for the grid normally used to calculate the energy and potential. We first consider some conditions that must be met in order to avoid inaccuracies in numerical calculations.

3.1. Conditions on a Grid. It would be possible, in principle, to work from the exact expression for the energy and take a suitable discretization on the grid. However, there is no a priori guarantee that the resulting forces would be consistent with the energy on the *same* grid. Problems with structural relaxation and energy conservation in molecular dynamics can arise if the force is not consistent with the discretized energy. We formulate two principles that must be followed in the derivation:

(1) The exchange-correlation potential must be the exact derivative of the approximate exchange-correlation energy. The Kohn−Sham equations are Euler equations that result from the condition of total energy stationarity with respect to the charge density. On a grid, the total energy will not be exact, but the approximate equations must nevertheless remain stationary.

(2) Forces must be exact derivatives of the approximate energy. Of course, because the energy is calculated on a grid, the expression for the force will also be an approximation. However, to use forces in structural relaxations, there must be well-defined zero-force atomic positions.

To enforce these principles, one must first approximate E_{XC} on a grid and then obtain the potential and forces from the resulting expression, rather than representing V_{XC} and $\mathbf{F}_i$ on the grid directly. If possible, the same grid should be used to represent the charge density, the energy, the potential, and the forces on atoms.

3.2. White−Bird Approach. In the next three sections, we will assume that we work on a regular grid, with three-dimensional periodic boundary conditions, which allows us to calculate gradients efficiently using FFTs; the resulting formulas can be easily generalized to other schemes such as atom-centered quadrature.[30] We start by approximating the GGA exchange-correlation energy (eq 8) as a sum over grid points l

$$E_{\mathrm{xc}}[\{n_l\}] = \omega \sum_l f_{\mathrm{xc}}(n_l, \mathbf{g}_l) \tag{11}$$

with ω the volume per grid point, the value of the charge density at a grid point l written as $n(\mathbf{r}_l) = n_l$ and its gradient $\nabla n(\mathbf{r}_l) = \mathbf{g}(\mathbf{r}_l) = \mathbf{g}_l$.

The first quantity to be obtained from the energy is the exchange-correlation potential. This was done by White and Bird,[29] who noted that although direct discretization of the exact potential (eq 10) is possible in this case, it is still inconvenient, because a grid twice as fine as the minimal grid (the density grid) is necessary to avoid inaccuracies.

The White-Bird approach notes that derivatives on the minimal grid are nothing but linear transformations on the *same* grid. For example, the gradient of the density can be expressed as

$$\mathbf{g}_l = \frac{1}{N}\sum_{m,l'} i\mathbf{G}_m n_{l'} e^{i\mathbf{G}_m\cdot(\mathbf{r}_l - \mathbf{r}_{l'})} \tag{12}$$

where we write the reciprocal lattice vectors associated with the uniform grid as $\mathbf{G}_m$. This transformation uses two Fourier transforms to form the gradient.

The transformation can be written as

$$\mathbf{g}_l = \sum_{l'} \mathbf{e}_{l,l} n_{l'} \tag{13}$$

with $\mathbf{e}_{l,l'} = -\mathbf{e}_{l',l}$ given by

$$\mathbf{e}_{l,l'} = \frac{1}{N}\sum_m i\mathbf{G}_m e^{i\mathbf{G}_m\cdot(\mathbf{r}_l - \mathbf{r}_{l'})} \tag{14}$$

Principle 1 from section 3.1 requires that

$$\delta E_{\mathrm{xc}} = \sum_l \frac{\mathrm{d}E_{\mathrm{xc}}}{\mathrm{d}n_l}\delta n_l = \omega \sum_l V_{\mathrm{xc},l}\delta n_l \tag{15}$$

Equivalently, using local orbitals as basis functions, the energy (eq 11) is minimized by varying the coefficients of the local orbitals (eq 3), $u_{i\alpha}^n$, so that

$$\frac{\partial E_{\mathrm{xc}}}{\partial u_{i\alpha}^n} = \omega \sum_l \frac{\mathrm{d}f_{\mathrm{xc},l}}{\mathrm{d}n_l}\frac{\partial n_l}{\partial u_{i\alpha}^n} \tag{16}$$

Both eqs 15 and 16 imply that the exchange-correlation potential for GGA functionals of the form (eq 8) is the total derivative $V_{\mathbf{xc},l} = \omega^{-1}\mathrm{d}E_{\mathrm{XC}}/\mathrm{d}n_l = \mathrm{d}f_{\mathrm{XC}}/\mathrm{d}n_l$. Because the KS matrix elements are also sums on a grid

$$\langle\phi_{i\alpha}|V_{\mathrm{xc}}|\phi_{j\beta}\rangle = \omega \sum_l \phi_{i\alpha,l}V_{\mathrm{xc},l}\phi_{j\beta,l} \tag{17}$$

with $\phi_{i\alpha,l} = \phi_{i\alpha}(r_l)$, the exchange-correlation potential can be expressed as

$$V_{\mathrm{xc},l} = \frac{\partial f_{\mathrm{xc}}}{\partial n_l} + \sum_{l'} \frac{\partial f_{\mathrm{xc}}}{\partial \mathbf{g}_{l'}} \cdot \frac{\partial \mathbf{g}_{l'}}{\partial n_l} = \frac{\partial f_{\mathrm{xc}}}{\partial n_l} + \sum_{l'} \frac{\partial f_{\mathrm{xc}}}{\partial \mathbf{g}_{l'}} \cdot \mathbf{e}_{l',l} \tag{18}$$

Note that this is the discrete counterpart of the exact potential (eq 10), but it was obtained directly from the discretized energy and yields an expression on the same grid as the density.

Once the discretized potential has been found, the Harris−Foulkes double-counting XC correction term results from following the original derivation[3,4] but applied to the energies on the density grid. The result is simply

$$\Delta E_{\mathrm{xc}}[\{n_l\}] = \omega \sum_l (f_{\mathrm{xc}}(n_l^{\mathrm{in}}, \mathbf{g}_l) - n_l^{\mathrm{in}} V_{\mathrm{xc},l}^{\mathrm{in}}) \tag{19}$$

3.3. Expression of the GGA NSC Force. To correctly derive the force on a grid, we proceed from the approximate energies (eqs 11 and 19) and the expression for the NSC force (eq 7). Considering only the NSC XC GGA force, we find that on a grid the correct expression is

$$\mathbf{F}_i^{\mathrm{NSC,xc}} = -\omega \sum_l \delta n_l \nabla_i V_{\mathrm{xc},l} \tag{20}$$

This is the natural discretization of the last term in eq 7.

From eq 18, the gradient of the potential, $\nabla_i V_{\mathbf{xc},l}$, necessary to evaluate the force, is

$$\nabla_{ip} V_{\mathrm{xc},l} = \frac{\partial^2 f_{\mathrm{xc},l}}{\partial n_l^2} \nabla_{ip} n_l + \sum_q \frac{\partial^2 f_{\mathrm{xc},l}}{\partial n_l \partial g_{lq}} \nabla_{ip} g_{lq} + \sum_{l'q} \frac{\partial^2 f_{\mathrm{xc},l'}}{\partial g_{l'q} \partial n_{l'}} e^q_{l',l} \nabla_{ip} n_{l'} + \sum_{l'qr} \frac{\partial^2 f_{\mathrm{xc},l'}}{\partial g_{l'q} \partial g_{l'r}} e^q_{l',l} \nabla_{ip} g_{l'r} \tag{21}$$

where p, q, and r are Cartesian components. Alternatively, using the linear relationship (eq 13) between the density and its gradient

$$\nabla_{ip} V_{\mathrm{xc},l} = \frac{\partial^2 f_{\mathrm{xc},l}}{\partial n_l^2} \nabla_{ip} n_l + \sum_{l'q} \frac{\partial^2 f_{\mathrm{xc},l}}{\partial n_l \partial g_{lq}} e^q_{l,l'} \nabla_{ip} n_{l'} + \sum_{l'q} \frac{\partial^2 f_{\mathrm{xc},l'}}{\partial g_{l'q} \partial n_{l'}} e^q_{l',l} \nabla_{ip} n_{l'} + \sum_{l''l'qr} \frac{\partial^2 f_{\mathrm{xc},l'}}{\partial g_{l'q} \partial g_{l'r}} e^q_{l',l} e^r_{l',l''} \nabla_{ip} n_{l''} \tag{22}$$

We now insert eq 22 into eq 20, rearrange and define, for clarity, the following quantities:

$$L_l^{(1)} = \omega \delta n_l \frac{\partial^2 f_{\mathrm{xc},l}}{\partial n_l^2} \tag{23}$$

$$L_l^{(2)} = \omega \sum_{l'q} \delta n_{l'} \frac{\partial^2 f_{\mathrm{xc},l}}{\partial g_{lq} \partial n_l} e^q_{l,l'} \tag{24}$$

$$L_l^{(3)} = -\omega \sum_{l'q} \delta n_{l'} \frac{\partial^2 f_{\mathrm{xc},l'}}{\partial n_{l'} \partial g_{l'q}} e^q_{l,l'} \tag{25}$$

$$L_l^{(4)} = -\omega \sum_{l'r} M_{l'r} e^r_{l,l'} \tag{26}$$

with

$$M_{lr} = \sum_{l'q} \delta n_{l'} \frac{\partial^2 f_{\mathrm{xc},l}}{\partial g_{lq} \partial g_{lr}} e^q_{l,l'} \tag{27}$$

where we have used the symmetry of the transformation $\mathbf{e}_{l,l'} = -\mathbf{e}_{l',l}$. $L_l^{(1)}$ is just a scalar quantity and $L_l^{(2)}$ is the dot product of a vector (a derivative of f_{xc}) and $\nabla \delta n_l$. Both $L_l^{(3)}$ and $L_l^{(4)}$ are divergences. Finally, M_{lr} is the dot product of $\nabla \delta n_l$ and a dyadic, that is, a vector.

The final expression for the non-self-consistent exchange-correlation force is simply

$$\mathbf{F}_i^{\mathrm{NSC,xc}} = -\sum_l L_l^{\mathrm{tot}} \nabla_i n_l = -\sum_l L_l^{\mathrm{tot}} \nabla_i \eta_i(|\mathbf{r}_l - \mathbf{R}_i|) \tag{28}$$

with $L_l^{\mathrm{tot}} = L_l^{(1)} + L_l^{(2)} + L_l^{(3)} + L_l^{(4)}$, that is, the sum of eqs 23−26, and $\mathbf{R}_i$ an atomic position, as usual. The second equality is true for superpositions of atomic densities η_i of the form of eq 4.

The continuum equivalent of the force is

$$\mathbf{F}_i^{\mathrm{NSC,xc}}(\mathbf{r}) = -\int \mathrm{d}\mathbf{r} L^{\mathrm{tot}}(\mathbf{r}) \nabla_i \eta_i(|\mathbf{r} - R_i|) \tag{29}$$

which happens to have the same basic dependence on the atomic densities as the LDA force,[5] but with a more complicated expression for the factor $L^{\mathrm{tot}}(\mathbf{r})$

$$L^{\mathrm{tot}}(\mathbf{r}) = \frac{\partial^2 f_{\mathrm{xc}}}{\partial n(\mathbf{r})^2} \delta n(\mathbf{r}) + \frac{\partial^2 f_{\mathrm{xc}}}{\partial n(\mathbf{r}) \partial \mathbf{g}(\mathbf{r})} \cdot \nabla \delta n(\mathbf{r}) - \nabla \cdot \left(\frac{\partial^2 f_{\mathrm{xc}}}{\partial n(\mathbf{r}) \partial \mathbf{g}(\mathbf{r})} \delta n(\mathbf{r}) + \frac{\partial^2 f_{\mathrm{xc}}}{\partial \mathbf{g}(\mathbf{r}) \partial \mathbf{g}(\mathbf{r})} \cdot \nabla \delta n(\mathbf{r}) \right) \tag{30}$$

This form can be used to derive discretizations for other integration grid schemes where necessary.

3.4. Implementation details. A subtlety about the implementation of eq 28 is that, in many widespread GGA functionals, the kernel frequently depends on the gradient only through its modulus and, therefore, all derivatives with respect to the gradient have to be expressed accordingly. The two relevant derivatives are the vector

$$\frac{\partial^2 f_{\mathrm{xc}}}{\partial n_l \partial \mathbf{g}_l} = \frac{\partial^2 f_{\mathrm{xc}}}{\partial n_l \partial |\mathbf{g}_l|} \frac{\mathbf{g}_l}{|\mathbf{g}_l|} \tag{31}$$

and a dyadic made of the following components:

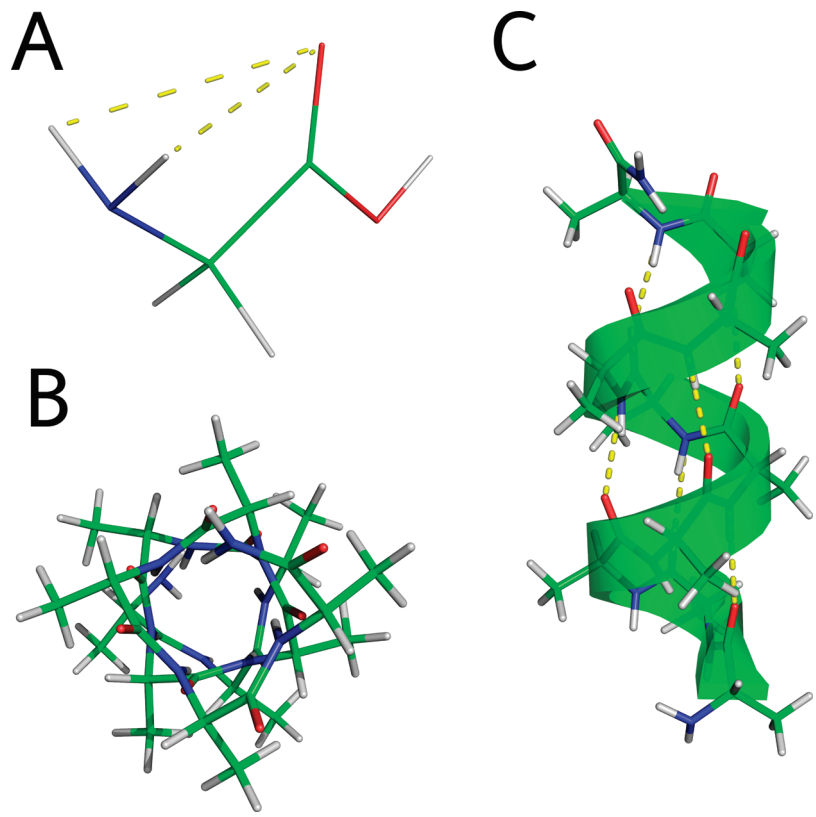
$$\frac{\partial^2 f_{\mathrm{xc}}}{\partial g_{lp} \partial g_{lq}} = \frac{\partial^2 f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|^2} \frac{g_{lp} g_{lq}}{|\mathbf{g}_l|^2} + \delta_{pq} \frac{\partial f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|} \frac{1}{|\mathbf{g}_l|} - \frac{\partial f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|} \frac{g_{lp} g_{lq}}{|\mathbf{g}_l|^3} \tag{32}$$

The latter appears in M_{lr}, in a dot product with $\nabla \delta n_l$, which can be simplified for implementation as

$$\frac{\mathbf{g}_l \cdot \nabla \delta n_l}{|\mathbf{g}_l|^2} \left(\frac{\partial^2 f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|^2} - \frac{1}{|\mathbf{g}_l|} \frac{\partial f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|} \right) \mathbf{g}_l + \frac{1}{|\mathbf{g}_l|} \frac{\partial f_{\mathrm{xc}}}{\partial |\mathbf{g}_l|} \nabla \delta n_l \tag{33}$$

When working on a regular grid, as we assume, it is possible to use only eight Fourier transforms to evaluate the force: four to determine the vector quantity $\nabla_i \delta n(\mathbf{r})$, which appears in eqs 24 and 27 (one direct transform of the scalar $\delta n(\mathbf{r})$, a product with the reciprocal lattice vectors and one inverse per component), and four to obtain the divergence of terms eqs 25 and 26, which can be computed together. This number of FFTs is identical to the computational effort needed to evaluate the functional itself, since eq 10, in the

Scheme 1. Self-Consistently Relaxed Structures of Glycine (A) and Amide-Terminated Deca-alanine in α-Helical Conformation, Viewed along Its Axis (B) and Laterally (C)[a]



[a] Hydrogen bonds are represented by dashed lines. Color scheme: C, green; N, blue; O, red; H, white.

White−Bird formulation, also involves the calculation of only one gradient and one divergence.

4. Applications

In this section, we will perform a NSC pre-relaxation before refining the structure fully self-consistently. GGA functionals are a useful test as they are more expensive to evaluate than LDA, and hence the potential savings are also more important.

We report on structure relaxations of glycine and two homopolymers of alanine, penta-alanine (5 amino acid residues) and deca-alanine (10 residues), using the PBE GGA functional.[31] We chose these systems because their relaxed structures strongly depend on hydrogen-bond interactions (Scheme 1). The structure of glycine was built de novo, but with a geometry close to the minimum energy conformer, that is, with the hydrogens of the amino group facing the carbonyl oxygen of the carboxylic acid. In this conformation, two hydrogen bonds are formed which contribute to stabilize the molecule. Both oligopeptides of alanine were built in an α-helix conformation, using an experimental structure (PDB 2DPQ) as a template.[32] This structure was obtained using X-ray diffraction crystallography, and hence, the hydrogens were missing. Furthermore, most residues were not alanines, and we built our model system by replacing all amino acid side chains by methyl groups, in arbitrary conformations.

Structural relaxations were carried out using a standard conjugate gradient (CG) algorithm in the implementation available in Conquest. The code uses periodic boundary conditions, but we made sure that the (orthorhombic) cell was big enough to prevent interactions between different images of the molecules. To avoid unwanted rotations, one bond of glycine was kept constrained to a fixed axis. In the polypeptides, only the side chains and hydrogens were relaxed: since the positions of the backbone atoms are experimental, in many practical situations we will want to keep them fixed even if they do not fully agree with our level of theory. The systems were small enough to make exact diagonalization of the Hamiltonian feasible. Consequently, the results should be of general validity and comparable to those obtained by other codes that use standard cubic scaling DFT. We note that the force formulas are identical for linear scaling, and have been used to test larger systems, such as DNA.[12]

All structures were relaxed by two methods. In the first one, a two-step protocol was applied. The initial structure was first relaxed non-self-consistently. The resultant intermediate structure was then relaxed fully self-consistently. We refer to this as method A. In the second method, the initial structure was completely relaxed self-consistently until convergence. This is method B.

Each CG step involves several evaluations of the functional, for two reasons. A line search is necessary to locate the minimum energy along the direction of the gradient. This search is common to both NSC and SCF methods. In addition, in SCF calculations, repeated evaluations are

Table 1. Comparison of SCF Relaxations of One Amino Acid and Two Peptides, with (Method A) and without (Method B) NSC Pre-relaxation[a]

		method A		method B	
		step 1 (NSC)	step 2 (SCF)	total 1 + 2	only SCF
glycine[b]	CG steps	32	16	48	21
	functional evaluations	116	733	849	805
	total energy (Ha)	−56.625449	−56.475861		−56.475870
penta-alanine[c]	CG steps	105	23	128	53
	functional evaluations	373	1180	1553	2961
	total energy (Ha)	−242.121314	−241.541173		−241.541182
deca-alanine[d]	CG steps	73	28	101	75
	functional evaluations	299	1081	1380	3036
	total energy (Ha)	−474.779129	−473.372474		−473.372670

[a] All relaxations were conducted using a conjugate gradient algorithm until the maximum force component was below 5.0×10^{-4} Ha Bohr^{-1}. NSC calculations were continued self-consistently in the second step of method A. [b] Ten atoms, two of them constrained to a fixed axis to avoid rotations [c] Fifty-four atoms, of which 38 were allowed to move and the rest (the heavy atoms of the backbone) were kept static [d] One hundred four atoms, of which only 73 moved.

Table 2. Root Mean Square Deviations (Å) between the Structures of the Molecules in Table *1* at Several Levels of Relaxation[a]

			method A		method B
			step 1 (NSC)	step 2 (SCF)	only SCF
glycine	unrelaxed		0.097	0.069	0.069
	method A	step 1 (NSC)		0.063	0.062
		step 2 (SCF)			0.008
penta-alanine	unrelaxed		0.279	0.269	0.258
	method A	step 1 (NSC)		0.082	0.094
		step 2 (SCF)			0.023
deca-alanine	unrelaxed		0.320	0.301	0.298
	method A	step 1 (NSC)		0.068	0.076
		step 2 (SCF)			0.017

[a] For each molecule, the compared structures are initial guess (unrelaxed), relaxed non-self-consistently (method A, step 1) and then self-consistently (method A, step 2), and relaxed fully self-consistently from the unrelaxed structure until convergence (Method B). The completely relaxed structures are labelled method A, step 2 (SCF) and method B, only SCF, and the RMSDs show that both methods lead essentially to the same minimum, in agreement with the energies in Table 1.

performed to reach self-consistency (typically, of the order of 10), while only one is necessary in the NSC case. Counting the total number of functional evaluations summed over all relaxation steps is, therefore, a better estimate of computational expense and calculation time than using the number of relaxation steps as a measure.

Table 1 compiles our results for the systems described above. While the total number of CG steps tends to increase if method A is applied, compared to a full SCF relaxation (method B), the NSC pre-relaxation consistently reduces the number of CG steps in the SCF part of the calculation (compare Step 2 of Method A with Method B in the table). This result suggests that the NSC equilibrium point is not far from the SCF one. Root mean square deviations (rmsd; Table 2) support this idea, with distances to the initial unrelaxed structure of similar magnitude for all (totally or partially) relaxed structures and RMSDs about 1 order of magnitude lower between the final relaxed structures by both methods. Thus, the rmsd values show that, at least in these examples, both methods lead to essentially the same energy minimum; this is an indication that problems with local minima are unlikely to be more common with NSC pre-relaxation. The total energies of the completely relaxed structures, also shown in Table 1, confirm this conclusion: the maximum difference found between the SCF values by both methods, which happened for deca-alanine, was 0.0002 Ha (or 0.1255 kcal/mol) and would satisfy chemical accuracy criteria.

Table 1 also shows that, except for glycine (a very small system), where the number of evaluations of the energy functional is very similar with and without NSC pre-relaxation, the NSC pre-relaxation can save a significant portion of energy evaluations. Indeed, this portion is as high as about 50% for both example peptides.

5. Discussion and Conclusions

This work presents the expression for the non-self-consistent exchange-correlation contribution to the DFT force on atoms when GGA functionals of the standard form are used. This expression is necessary to perform non-self-consistent ab initio tight binding electronic structure calculations with this important class of functionals. It also makes it possible to carry out NSC pre-relaxations of condensed-matter systems and molecular structures. Moreover, the expression of the force can also be used to correct the forces of poorly converged self-consistent calculations.

We have shown that the expression can be computed on a regular grid using only 8 Fourier transforms. This is the same as for the evaluation of the functional itself, which makes the force calculation efficient. Furthermore, as was the case for LDA functionals, the expression remains

unchanged in linear scaling algorithms and can thus be applied to large systems.

We have presented geometry optimization results on small and medium size biomolecular systems using the Conquest code. Although we have not used the linear-scaling capabilities of the code in this work, large biomolecules, in particular hydrated DNA,[12] have already been studied with Conquest. The calculations in this work show that NSC pre-relaxations can assist in finding equilibrium geometries faster than by using only self-consistent methods. This will be important in larger problems, for which $O(N)$ will be essential.

For the two small alanine peptides considered, there were important savings (of almost 50%) in the total number of functional evaluations, the most expensive part of the calculation. (For glycine, the smallest amino acid, the pre-relaxation made only a small difference: 849 and 805 Harris−Foulkes energy evaluations respectively). This is a promising result and indicates that the method could be of general validity. However, to gain confidence in this conclusion, more calculations are needed. In particular, globular proteins could be more difficult problems than the simple α-helical structures presented here. Finally, we should note that we used a conjugate gradient algorithm. The inter-relationships of the pre-relaxation method with other, more efficient geometry optimization approaches should be evaluated in the future.

Acknowledgment. A.S.T. was a Ramón Areces post-doctoral fellow and is funded by ICYS-IMAT NIMS. D.R.B. is supported by the Royal Society. This work is partly supported by Grant-in-Aid for Scientific Research (KAKENHI) from the MEXT and JSPS, Japan. Calculations where performed in the Theory Cluster at LCN, UCL, London, and in the Numerical Materials Simulator at NIMS, Tsukuba, Japan.

Supporting Information Available: Coordinates of glycine, penta-alanine, and deca-alanine, before and after relaxation by the two methods described, have been deposited. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Himo, F.; Siegbahn, P. E. M. *Chem. Rev.* **2003**, *103*, 2421–2456.

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

(3) Harris, J. *Phys. Rev. B* **1985**, *31*, 1770–1779.

(4) Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B* **1989**, *39*, 12520–12536.

(5) Miyazaki, T.; Bowler, D. R.; Choudhury, R.; Gillan, M. J. *J. Chem. Phys.* **2004**, *121*, 6186–6194.

(6) Bowler, D. R.; Miyazaki, T.; Gillan, M. J. *J. Phys.: Condens. Matter.* **2002**, *14*, 2781–2798.

(7) Bowler, D. R.; Choudhury, R.; Gillan, M. J.; Miyazaki, T. *Phys. Status Solidi B* **2006**, *243*, 989–1000.

(8) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

(9) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.

(10) Heady, L.; Fernandez-Serra, M.; Mancera, R. L.; Joyce, S.; Venkitaraman, A. R.; Artacho, E.; Skylaris, C.-K.; Ciacchi, L. C.; Payne, M. C. *J. Med. Chem.* **2006**, *48*, 5141–5153.

(11) Gillan, M. J.; Bowler, D. R.; Torralba, A. S.; Miyazaki, T. *Comput. Phys. Commun.* **2007**, *177*, 14–18.

(12) Otsuka, T.; Miyazaki, T.; Ohno, T.; Bowler, D. R.; Gillan, M. J. *J. Phys.: Condens. Matter* **2008**, *20*, 294201.

(13) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726–2734.

(14) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274–7280.

(15) Hübsch, A.; Endres, R. G.; Cox, D. L.; Singh, R. R. P. *Phys. Rev. Lett.* **2005**, *94*, 178102.

(16) Kaschner, R.; Hohl, D. *J. Phys. Chem. A* **1998**, *102*, 5111–5116.

(17) Bowler, D. R.; Gillan, M. J. *Chem. Phys. Lett.* **2002**, *355*, 306–310.

(18) Kenny, S. D.; Horsfield, A. P.; Fujitani, H. *Phys. Rev. B* **2000**, *62*, 4899–4905.

(19) Ozaki, T. *Phys. Rev. B* **2003**, *67*, 155108.

(20) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745–2779.

(21) Torralba, A. S.; Todorović, M.; Brázdová, V.; Choudhury, R.; Miyazaki, T.; Gillan, M. J.; Bowler, D. R. *J. Phys.: Condens. Matter* **2008**, *20*, 294206.

(22) Hernández, E.; Gillan, M. J.; Goringe, C. M. *Phys. Rev. B* **1997**, *55*, 13485–13493.

(23) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *J. Chem. Phys.* **2005**, *122*, 084199.

(24) Sankey, O. F.; Niklewski, D. J. *Phys. Rev. B* **1989**, *40*, 3979–3995.

(25) Bendt, P.; Zunger, A. *Phys. Rev. Lett.* **1983**, *50*, 1684–1688.

(26) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(27) Anglada, E.; Junquera, J.; Soler, J. M. *Phys. Rev. E* **2003**, *68*, 055701.

(28) Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168–3171.

(29) White, J. A.; Bird, D. M. *Phys. Rev. B* **1994**, *50*, 4954–4957.

(30) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547–2553.

(31) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(32) Cnudde, S. E.; Prorok, M.; Dai, Q.; Castellino, F. J.; Geiger, J. H. *J. Am. Chem. Soc.* **2007**, *129*, 1586–1593.

CT8005425

JCTC Journal of Chemical Theory and Computation

Addressing Through-H Magnetic Interactions: A Comprehensive ab Initio Analysis of This Efficient Coupler

Boris Le Guennic,[†] Nadia Ben Amor,[‡,§] Daniel Maynau,[‡,§] and Vincent Robert*[,†]

Université de Lyon, Ecole Normale Supérieure de Lyon and CNRS, Laboratoire de Chimie, 46 allée d'Italie, 69364 Lyon Cedex 07, France, Université de Toulouse, UPS, LCPQ (Laboratoire de Chimie et Physique Quantiques), IRSAMC, 118 route de Narbonne, 31062 Toulouse cedex, France, and CNRS, Université de Toulouse, UPS, LCPQ (Laboratoire de Chimie et Physique Quantiques), IRSAMC, 118 route de Narbonne, 31062 Toulouse cedex, France

Received December 9, 2008

Abstract: The exchange coupling in a structuraly characterized Cu^{II}_2 complex is analyzed to highlight the role of H bonds in the generation of efficient magnetic interactions. The interest for complementary insights which are not accessible through DFT calculations (Desplanches, C. et al. *J. Am. Chem. Soc.* **2002**, *124*, 5197) has driven this state-of-the-art ab initio inspection. The wave function expansion based upon localized orbitals allows us to selectively turn on specific mechanisms and quantitatively evaluate their roles in the exchange interactions. Our singlet−triplet splitting calculations demonstrate the enhancement of the magnetic coupling through a concerted oxygen-to-metal charge transfer and electronic redistribution within the OH bond of the OH···O magnetic linker. This mechanism accounts for ∼35% of the total experimentally measured singlet−triplet energy difference. This analysis strongly suggests that H bonds might be particularly useful not only in the establishment of intermolecular contacts but also within the basic units of magnetic materials.

Introduction

The importance of hydrogen bonds[1] in biology, physics, and chemistry has been much debated since these weak bonds might be at the origin of fundamental phenomena such as DNA structuration, biochemical reactions, and spin transition. While the former phenomena have been evidenced and much studied for several decades, it is more recently that magnetic systems involving H bonds have been the subject of intense research due to the tremendously promising spin-crossover behavior.[2] The possibility to generate bistability using various nuclearities building blocks interacting through weak bonds contacts has been suggested as an original synthetic route.[3] Cooperativity is known to be of particular importance, and its origin might be found in π-stacking or H-bond networks formation. The quantification of such weak bonds using quantum chemical calculations has thus become a challenging issue.

Traditionally, the analysis of magnetically coupled systems relies on the Heisenberg Hamiltonian H assuming that the spatial parts of the spin-states wave functions are very similar. Thus, one introduces a so-called exchange coupling constant J which is expected to depict the low-energy spectroscopy writing $H = -JS_1S_2$. For a dinuclear species with one unpaired electron on each metal ion, the resulting singlet−triplet energy separation $E_T - E_S$ simply reads J. The corresponding value extracted from magnetic susceptibility measurements can be compared to quantum chemical calculations.

Even though density functional theory (DFT) based methods have often reached good agreement in the determination of such constants,[4] they may not give access to

* Corresponding author E-mail: vincent.robert@ens-lyon.fr.
[†] Université de Lyon.
[‡] Université de Toulouse.
[§] CNRS, Université de Toulouse.

10.1021/ct900022k CCC: $40.75
Published on Web 05/04/2009

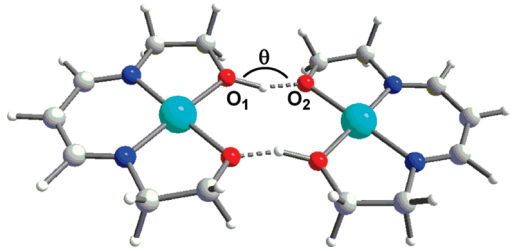


Figure 1. Cu_2 centrosymmetric model complex **1**. The experimental $O_1H\cdots O_2$ angle θ is 159°.

the underlying mechanisms which govern the spin-states ordering. Besides, it is known that the evaluation of weak interactions relies on explicitly correlated ab initio calculations which are flexible enough to account, for instance, for the dynamical charge fluctuations (i.e., dipole−dipole interactions). Nevertheless, their accurate evaluation remains a challenging issue of constant interest since the system sizes preclude the use of such quantum theory methods. This kind of wave-function-based calculation has proved to reach spectroscopic accuracy for magnetically coupled extended materials.[5]

We report herein a detailed investigation of the exchange interactions in the Cu_2 model complex **1** (see Figure 1), extracted from the reported $[Cu(DiimH)]_2$,[6] where four methyl groups were replaced by H atoms. The theoretical strategy we suggest allows us to (i) perform ab initio wave function calculations (i.e., configurations interaction, CI) on these rather extended architectures and (ii) quantify the contributions arising from the hydrogen-bond backbone. One may wonder how much the $O_1H\cdots O_2$ (see Figure 1) structural and electronic characteristics influence the nature and amplitude of the magnetic interactions between the two paramagnetic centers. Part of the answer is to be found in quantum chemistry calculations in which specific mechanisms can be turned on at will. In that sense, this work complements the quantitative evaluations which are accessible by means of DFT-based approaches. In fact, the reading of the multideterminantal wave functions based upon localized orbitals (i.e., valence-bond-like picture) offers a step-by-step analysis of the relevant mechanisms accounting for the spin-states hierachization. By selecting the valence orbitals, one can grasp their role in the establishment of efficient magnetic channels. Particular attention is paid to the $O_1H\cdots O_2$ bridge geometry and the valence orbitals involved in the exchange mechanism. A detailed analysis of the correlated wave function suggests that the charge fluctuation within the OH bond facilitates the ligand-to-metal charge transfer (LMCT). This superexchange-like mechanism is very sensitive to the $O_1H\cdots O_2$ angle θ since colinearity along the H bond greatly enhances (∼20%) the antiferromagnetic behavior.

Computational Details

The coupling between the magnetic moments localized on the Cu^{II} metal ions was investigated using wave-function-based calculations. This framework is particularly appealing in the microscopic analysis of the exchange interactions. With this goal in mind, the difference dedicated configurations interaction (DDCI) method[7] has been designed and successfully applied[8] to evaluate vertical energy differences. First, complete active space self-consistent field (CASSCF) calculations were performed to generate a reference space including the leading electronic configurations in the desired spin multiplicities, singlet ($S = 0$, S) and triplet ($S = 1$, T). The minimal active space includes two electrons localized upon the two mainly d-type molecular orbitals (CAS[2,2]). In order to grasp the importance of the hydrogen bonds in the exchange coupling, this minimal active space was enlarged to include the $O_1H\cdots O_2$ bridge molecular orbitals (MOs) in the reference wave function. Thus, a CAS[10,8] (10 electrons in 8 MOs) was considered to account for the contributions arising from the oxygen atoms lone pairs (n_{O2}) and bonding and antibonding O_1H groups MOs (σ_{O1H} and σ^*_{O1H}). This particular CAS is flexible enough to incorporate all the contributions arising from the charge fluctuations within the $O_1H\cdots O_2$ bridges. All CASSCF calculations were performed using the Molcas 6.0 package[9] and available ANO-RCC-type functions. The Cu atoms were described with a (21s15p10d6f4g2h)/[5s4p3d] contraction.[10] Particular attention was paid to the bridge parts and first nearest neighbors of the metal ions. Thus, a (14s9p4d3f2g)/[3s2p1d] contraction was used for the N and C atoms, whereas the O atoms were described by a (14s9p4d3f2g)/[3s3p1d] contraction.[11] A (8s4p3d1f)/[2s1p] contraction was used for the H atoms involved in the hydrogen bonds.[12] The other H atoms were depicted with a smaller basis set (8s4p3d1f)/[2s]. In a valence-bond (VB) type framework, orthogonal localized orbitals (LOs) were finally constructed using the canonical CASSCF orbitals. These LOs allow for a chemically intuitive analysis of the relevant mechanisms accompanying the singlet−triplet hierarchization.

The dynamical polarization and correlation effects were then incorporated using the DDCI method as implemented in the CASDI code.[13] A detailed analysis of the underlying mechanisms was given in the pioneer work of de Loth et al.[14] It has been clearly demonstrated that a bare valence-only description (i.e., CASCI energy difference) is not reliable to accurately grasp energy differences as small as a few tens of wavenumbers.[15] Thus, one should include selected configurations reached by excitations on top of the CASSCF wave function. As the number of degrees of freedom (i.e., holes in doubly occupied (inactive) MOs or particles generated in empty (virtual) MOs) grows, the successive CAS+S (also referred to as DDCI-1), CAS+DDCI-2, and CAS+DDCI-3 levels are reached by expanding the CI space. Since the DDCI philosophy relies on the simultaneous characterization of different states which share similar spatial descriptions, a common set of MOs must be initially determined to build up the CI space. The CASSCF triplet LOs were used. Nevertheless, we checked that these J values are almost unaffected by the use of the singlet state LOs, whatever the level of calculations.

Following the minimal active space CAS[2,2] inspection, we were able to reach a DDCI-3 level of calculations. This is to be contrasted with the use of the enlarged CAS[10,8] which disposes any calculation beyond CAS+S. Neverthe-

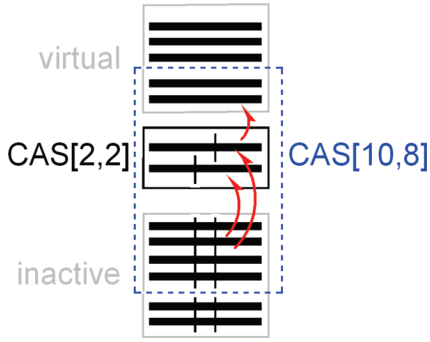


Figure 2. Most important CAS[2,2]+DDCI-3 mechanisms included within the CAS[10,8] strategy.

less, it has been shown that inclusion of the bridge MOs allows one to reach spectroscopic accuracy at a CAS+S level.[16] Indeed, this enlarged CAS[10,8] incorporates all the electronic configurations which are likely to qualitatively discriminate the singlet from the triplet states. In particular, it contains all the leading CAS[2,2]+DDCI-3 mechanisms, i.e., the ones which involve the bridges MOs (see Figure 2). At a CAS[10,8]+S level, the important instantaneous charge relocalization upon the external part of the ligands are taken into account.[14] Besides, some particular mechanisms involving the charge redistributions within the O_1H bonds are explicitly turned on ($\sigma_{O1H}^2\ \sigma^*_{O1H}{}^0 \rightarrow \sigma_{O1H}{}^0\ \sigma^*_{O1H}{}^2$ double excitation process), and their coupling to LMCT can be evaluated. This concerted scenario, which allows for the electron transfer from one Cu^{II} center to the other, is absent in the CAS[2,2]-based DDCI-3 calculations.

Complementary DFT calculations were performed to clarify the preference for linearity in the $O_1H\cdots O_2$ fragment. We used the B3LYP exchange-correlation functional with triple-ζ basis sets as implemented in the Gaussian03 code.[17] Starting from the triplet DFT MOs, the broken-symmetry (BS) state[18] was converged to extract the exchange coupling constant $J = E_{BS} - E_T$. Even though the J extraction based upon unrestricted DFT calculations is still controversial,[4a,19] we used the same strategy as suggested in ref 20.

Results and Discussion

On the basis of quantitative DFT calculations,[20] it has been previously stated that the H bridges play essentially an indirect structural role. Nevertheless, the single-reference character of the DFT wave function does not provide a microscopic picture for the underlying mechanisms. In particular, how much a H bond gets involved in the establishement of efficient magnetic interaction is of particular importance in the preparation of magnetic materials. To complement the DFT data and offer a detailed analysis of the magnetic exchange mechanism, we first performed multireference CASSCF calculations on **1**. Considering the d^9 electronic configuration of the Cu^{II} ion, the minimal active space consists of two electrons in two molecular orbitals (MOs). As expected, the active MOs of both S and T states are essentially the in-phase and out-of-phase linear combinations of the Cu d_{xz} atomic orbitals (see Figure 3).

The CAS[2,2]SCF energy difference $\Delta E = E_S - E_T$ leads to a poor estimation of $J \approx -9$ cm^{-1}. Nevertheless,

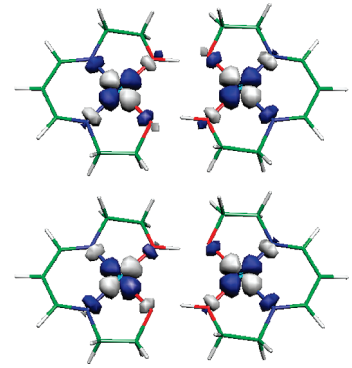


Figure 3. Magnetic MOs a_g (left) and a_u (right) extracted from a CAS[2,2] calculation over the T state.

Table 1. Calculated Exchange-Coupling Constant J (cm^{-1}) in Complexes **1** and **2**[a]

	minimal valence space, CAS[2,2]+DDCI-3	extended valence space, CAS[10,8]+S	DFT
1	−62	−106	−100
2	−81	−128	−119

[a] $J_{exp} = -94$ cm^{-1}.

subsequent CI calculations significantly changed this state of affairs. Along the reported DDCI scheme,[7] one includes a selection of determinants constructed on the triplet MOs. Since these calculations might be out of reach due to the large CI space, a strategy consists of building LOs to take advantage of the local character of the correlation effects. This valence-bond-type (VB) description allows one to concentrate on the most relevant determinants.[21]

From our DDCI-3 calculations, the $S-T$ energy difference J is found to be −62 cm^{-1} (Table 1), reflecting a ∼35% deviation from the reported experimental value −94 cm^{-1}.[6] This numerical discrepancy is rather puzzling since the DDCI framework offers spectroscopic accuracy in the investigation of magnetically coupled systems.[8] Therefore, this non-negligible deviation from both experimental[6] and previous DFT estimation (−87 cm^{-1})[20] suggests that either (i) the H atomic positions are ill defined since an intuitive chemical picture would anticipate quasi-linearity of the $O_1H\cdots O_2$ bridge or (ii) some important mechanisms contributing to the exchange coupling scheme are missing in the minimal valence space approach.

Thus, DFT/UB3LYP and similar CAS[2,2]+DDCI-3 calculations were performed upon a hypothetical structure **2** constructed from **1** by setting $\theta = 180°$. Along this deformation, we maintained the O−H bond distances ratio constant, imposing a simultaneous shortening (∼0.02 Å) of these particular distances. As seen in Table 1, $|J|$ is significantly enhanced as the $O_1H\cdots O_2$ angle θ reaches 180°, in agreement with previous calculations.[20] A similar conclusion holds in light of the DDCI-3 calculations which exhibit a −81 cm^{-1} exchange constant in **2**. These results demonstrate the apparent role of the H bonds in the establishment of efficient exchange coupling channels. At this stage, let us mention that the ground state S is lowered by ∼10 000 cm^{-1} as the geometry is modified from **1** to **2**. This energy stabilization was confirmed by full geometry DFT/UB3LYP optimization which highlights a quasi-linearity of the

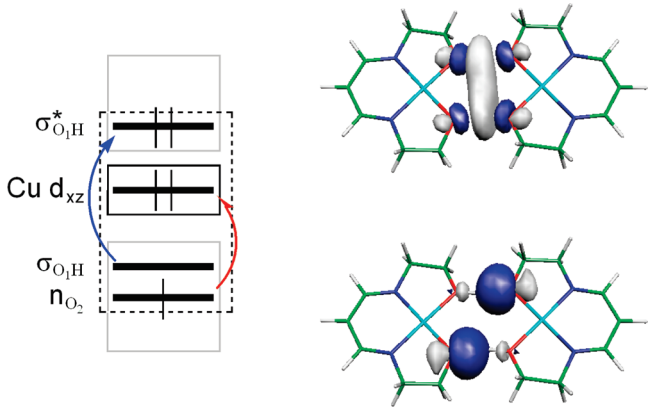


Figure 4. Selected active spaces (a_g symmetry): minimal valence space (CAS[2,2], full-line box) and extended valence space (CAS[10,8], dashed-line box). The blue arrow stands for a double excitation, whereas the red harpoon corresponds to a single excitation. The n_{O2} (bottom) and σ^*_{O1H} (top) orbitals are sketched on the right-hand side.

$O_1H\cdots O_2$ bridge ($\theta = 178°$) while the rest of the structure is negligibly modified. Even though the crystallization conditions are likely to control the molecular units structuration, our "magneto-structural" inspection suggests that the role of the hydrogen bonds may not be limited to a structural one.[20] The oxygen atom lone pairs n_{O2} might be directly involved in a superexchange-like mechanism with the adjacent O_1H group σ_{O1H} and σ^*_{O1H} orbitals, making the $O_1H\cdots O_2$ fragment an efficient coupler.

Thus, the active space was enlarged to CAS[10,8] to incorporate the valence orbitals and electrons of the $O_1H\cdots O_2$ linkers (see Figure 4). In a CAS[2,2] approach, the kinetic exchange[22] is explicitly introduced. At a DDCI-3 level, this contribution is not only revisited but the mechanisms involving the bridging ligand orbitals are turned on.[14] However, a mechanism involving the simultaneous charge reorganization within the OH bond and oxygen-to-metal charge transfer is absent and might play a determinant role. Thus, the CAS[10,8] strategy allows one to evaluate the contributions of superexchange mechanisms which were not accessible within the minimal CAS[2,2] approach. While the CAS[10,8]SCF S and T energies are easily obtained, the required CI treatment is much more demanding. Nevertheless, the VB-type description based on the LOs considerably reduces the CI space. Besides, the use of this particular active space including the valence LOs of the bridge (i.e., n_{O2}, σ_{O1H}, σ^*_{O1H}) affords a limited expansion of the wave function to single excitations (so-called CAS+S space).[16] As seen in Table 1, our CAS[10,8]+S result displays quantitative agreement (deviation smaller than 12%) with the experimental data since J is calculated to be -106 cm^{-1}. Besides, the reading of the wave function based on the bond LOs sheds the light on the importance of some specific electronic configurations (i.e., charge transfer forms) which may participate in the exchange coupling interaction. Any mechanism which enhances the effective hopping integral between the two Cu^{II} ions is likely to favor the antiferromagnetic behavior.[23] Since the antibonding orbital σ^*_{O1H} is mostly localized upon the H atom, the $\sigma_{O1H}{}^2\sigma^*_{O1H}{}^0 \rightarrow \sigma_{O1H}{}^0\sigma^*_{O1H}{}^2$ double-excitation process (see Figure 4) tends to accumulate electronic density upon the H atom and deplete the O_1 one. Interestingly, this instantaneous charge reorganization within the O_1H bonds turned out to be concerted with the ligand-to-metal $n_{O2} \rightarrow d_{xz}$ excitations (see Figure 4). Finally, the corresponding electronic configuration amplitude, though small, is increased by an order of magnitude when the $O_1H\cdots O_2$ fragment becomes linear. The comparison with the previous minimal valence space calculations is rather instructive. In fact, the here-evidenced electronic circulation was not included in the CAS[2,2] calculations and led to an underestimation of the effective hopping parameter. This is to be contrasted with the improved $S-T$ energy difference as soon as the intrinsically through-H bond superexchange mechanism is turned on. Finally, the corresponding electronic density fluctuations bring a complementary stabilization (47 cm^{-1} as compared to the minimal CAS description) of the S state over the T state in complex **2**. In light of the ~36% deviation with the reported experimental value, one can conclude that a much better agreement with experiment is reached for structure **1** and structure **2** might not be a satisfactory candidate. Besides, the need for an enlarged CAS[10,8] demonstrates the importance of charge fluctuations within the $O_1H\cdots O_2$ motif.

Conclusion

Our study sheds new light on to the prime role of hydrogen bonds in magnetically coupled systems and the strength of wave function methods to convey interpretative pictures. DFT-based calculations have demonstrated satisfactory accuracy in magnetic coupling evaluation, even if they remain biased due to the arbitrariness of the exchange-correlation functional. Moreover, they may not be well adapted to provide such pictures since they rely on single-determinant expansions of the wave functions. From our inspection using wave-function-based calculations, the deviation between experimental and a minimal valence space approach can be ascribed to a non-negligible electronic circulation within the $O_1H\cdots O_2$ bridge. This contribution, possibly among others, accounts for a supplementary 44 cm^{-1} stabilization of the singlet state over the triplet state. As expected, the coupling constant is very sensitive to the $O_1H\cdots O_2$ angle, and it is demonstrated that the hydrogen bond is directly involved in a superexchange-like mechanism. Using localized orbitals, we demonstrate that the H bond should not be a priori disposed when exchange coupling constants are considered in any magnetic systems. Our initial hypothesis regarding the crystallographic positions of the bridging hydrogen atoms can be discarded in light of comparative experimental and theoretical exchange constant values. The rational design of magnetic materials incorporating cooperativity effects through weak bond contacts should benefit from this theoretical inspection.

Acknowledgment. The authors thank the "Institut du Développement et des Ressources en Informatique Scientifique" (IDRIS) for computing facilities. This research was supported by the ANR (contract no. ANR-07-JCJC-0045-01) (*fdp*-magnets) project.

References

(1) Latimer, W. M.; Rodebush, W. H. *J. Am. Chem. Soc.* **1920**, *42*, 1419–1433.

(2) See, for instance: (a) Gütlich, P.; Jung, J.; Goodwin, H. A. Spin transition in iron(II) complexes. In *Molecular magnetism: from molecular assemblies to the devices*; Coronado, E. et al., Eds.; NATO Advance Study Institute Series E321; Plenum: New York, 1996; p 327. (b) Matouzenko, G. S.; Bousseksou, A.; Lecocq, S.; van Koningsbruggen, P. J.; Perrin, M.; Kahn, O.; Collet, A. *Inorg. Chem.* **1997**, *36*, 5869–5879.

(3) Kahn, O.; Martinez, C. J. *Science* **1998**, *279*, 44–48.

(4) See, for instance: (a) Ruiz, E.; Cano, J.; Alvarez, S.; Alemany, P. *J. Comput. Chem.* **1999**, *20*, 1391–1400. (b) Rudra, I.; Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *124*, 024103. (c) Cauchy, T.; Ruiz, E.; Alvarez, S. *J. Am. Chem. Soc.* **2006**, *128*, 15722–15727, and references therein.

(5) See, for instance: (a) de Graaf, C.; Moreira, I. de P. R.; Illas, F.; Martin, R. L. *Phys. Rev. B* **1999**, *60*, 3457–3464. (b) Munoz, D.; Illas, F.; Moreira, I. de P. R. *Phys. Rev. Lett.* **2000**, *84*, 1579–1582. (c) Moreira, I. de P. R.; Suaud, N.; Guihéry, N.; Malrieu, J.-P.; Caballol, R.; Bofill, J. M.; Illas, F. *Phys. Rev. B* **2002**, *66*, 134430. (d) Cabrero, J.; de Graaf, C.; Bordas, E.; Caballol, R.; Malrieu, J.-P. *Chem.—Eur. J.* **2003**, *9*, 2307–2315.

(6) Bertrand, J. A.; Black, T. D.; Eller, P. G.; Helm, F. T.; Mahmood, R. *Inorg. Chem.* **1976**, *15*, 2965–2970.

(7) (a) Miralles, J.; Daudey, J.-P.; Caballol, R. *Chem. Phys. Lett.* **1992**, *198*, 555–562. (b) Miralles, J.; Castell, O.; Caballol, R.; Malrieu, J.-P. *Chem. Phys.* **1993**, *172*, 33–43.

(8) See, for instance: (a) Herebian, D.; Wieghardt, K. E.; Neese, F. *J. Am. Chem. Soc.* **2003**, *125*, 10997–11005. (b) Messaoudi, S.; Robert, V.; Guihéry, N.; Maynau, D. *Inorg. Chem.* **2006**, *45*, 3212–3216. (c) Le Guennic, B.; Robert, V. *C. R. Chimie* **2008**, *11*, 650–664. (d) de Graaf, C.; Illas, F. *Phys. Rev. B* **2001**, *63*, 014404. (e) Suaud, N.; Lepetit, M.-B. *Phys. Rev. B* **2000**, *62*, 402–409.

(9) Karlström, G.; Lindh, R.; Malmqvist, P.-A.; 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.

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

(11) Roos, B. O.; Lindh, R.; Malmqvist, P.-A.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

(12) Widmark, P.-O.; Malmqvist, P.-A.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.

(13) Ben Amor, N.; Maynau, D. *Chem. Phys. Lett.* **1998**, *286*, 211–220.

(14) de Loth, P.; Cassoux, P.; Daudey, P.; Malrieu, J. P. *J. Am. Chem. Soc.* **1981**, *103*, 4007–4016.

(15) (a) Calzado, J. C.; Cabrero, J.; Malrieu, J. P.; Caballol, R. *J. Chem. Phys.* **2002**, *116*, 2728–2747. (b) Calzado, J. C.; Cabrero, J.; Malrieu, J. P.; Caballol, R. *J. Chem. Phys.* **2002**, *116*, 3985–4000.

(16) Gellé, A.; Munzarova, M. L.; Lepetit, M.-B.; Illas, F. *Phys. Rev. B* **2003**, *68*, 125103.

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

(18) (a) Noodleman, L.; Norman, J. G. *J. Chem. Phys.* **1979**, *70*, 4903. (b) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737. (c) Noodleman, L.; Case, D. A. *Adv. Inorg. Chem.* **1992**, *38*, 423. (d) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J.-M. *Coord. Chem. Rev.* **1995**, *144*, 199.

(19) (a) Ruiz, E.; Cano, J.; Alvarez, S.; Alemany, P. *J. Comput. Chem.* **1999**, *20*, 1391–1400. (b) Caballol, R.; Castell, O.; Illas, F.; Moreira, I. de P. R.; Malrieu, J. P. *J. Phys. Chem. A* **1997**, *101*, 7860–7866.

(20) Desplanches, C.; Ruiz, E.; Rodriguez-Fortea, A.; Alvarez, S. *J. Am. Chem. Soc.* **2002**, *124*, 5197–5205.

(21) (a) Rota, J.-B.; Norel, L.; Train, C.; Ben Amor, N.; Maynau, D.; Robert, V. *J. Am. Chem. Soc.* **2008**, *130*, 10380–10385. (b) Le Guennic, B.; Petit, S.; Chastanet, G.; Pilet, G.; Luneau, D.; Ben Amor, N.; Robert, V. *Inorg. Chem.* **2008**, *47*, 572–577. (c) Bories, B.; Maynau, D.; Bonnet, M.-L. *J. Comput. Chem.* **2007**, *28*, 632–643.

(22) Anderson, P. W. *Solid State Phys.* **1963**, *14*, 99–214.

(23) Kahn, O. *Molecular Magnetism*; VCH: New York, 1993.

CT900022K

JCTC Journal of Chemical Theory and Computation

Conformers of Gaseous Cysteine

Jeremiah J. Wilke,[†] Maria C. Lind,[†] Henry F. Schaefer III,[†] Attila G. Császár,*[,‡] and Wesley D. Allen*[,†]

Department of Chemistry and Center for Computational Chemistry, University of Georgia, Athens, Georgia 30602, and Laboratory of Molecular Spectroscopy, Institute of Chemistry, Eötvös University, H-1518 Budapest 112, P.O. Box 32, Hungary

Received January 3, 2009

Abstract: Structures, accurate relative energies, equilibrium and vibrationally averaged rotational constants, quartic and sextic centrifugal distortion constants, dipole moments, ^{14}N nuclear quadrupole coupling constants, anharmonic vibrational frequencies, and double-harmonic infrared intensities have been determined from ab initio electronic structure computations for conformers of the neutral form of the natural amino acid L-cysteine (Cys). A systematic scan located 71 unique conformers of Cys using the MP2(FC)/cc-pVTZ method. The large number of structurally diverse low-energy conformers of Cys necessitates the highest possible levels of electronic structure theory to determine their relative energies with some certainty. For this reason, we determined the relative energies of the lowest-energy eleven conformers, accurate within a standard error (1σ) of about 0.3 kJ mol^{-1}, through first-principles composite focal-point analyses (FPA), which employed extrapolations using basis sets as large as aug-cc-pV(5+d)Z and correlation treatments as extensive as CCSD(T). Three and eleven conformers of L-cysteine fall within a relative energy of 6 and 10 kJ mol^{-1}, respectively. The vibrationally averaged rotational constants computed in this study agree well with Fourier-transform microwave spectroscopy results. The effects determining the relative energies of the low-energy conformers of cysteine are analyzed in detail on the basis of hydrogen bond additivity schemes and natural bond orbital analysis.

1. Introduction

Because amino acids are the building blocks of proteins and peptides, the structural investigation of them, extending from solids to the gas phase, has received considerable experimental and theoretical attention.[1] Cysteine (Cys) is the only amino acid with a reactive sulfur moiety. In this regard, cysteine contributes to diverse structures, including disulfide bonds, zinc fingers, and Fe−S coordination in iron−sulfur proteins.[2] Functionally, disulfide bonds formed from cysteine serve a central role in glutathione, a mediator of oxidative stress, and strong nucleophilicity also makes cysteine a key component of the active site in many other enzymes.[3,4]

In the gas phase, amino acids are intrinsically flexible systems, occurring in their neutral form and exhibiting a large number of low-energy conformers. Even glycine, with only three rotatable single bonds, has eight conformers, five of which have relative energies less than 12 kJ mol^{-1}.[1,5,6] The number of natural amino acids is limited, and the relatively small size of these molecules allows the application of highly sophisticated quantum chemical methods to study their equilibrium and dynamical structures and rotational−vibrational spectra. The number of local minima on the respective potential energy surfaces (PES) and the structural properties of the related conformers, including accurate relative energies, are available for a number of amino acids. A review summarizing results before 1999 is provided in ref 1. Given the accuracy of modern electronic structure techniques, characterization of the complex PESs of amino acids should precede and supplement related experimental

* Corresponding author e-mail: wdallen@uga.edu (W.D.A.); csaszar@chem.elte.hu (A.G.C.).

[†] University of Georgia.

[‡] Eötvös University.

10.1021/ct900005c CCC: $40.75
Published on Web 05/15/2009

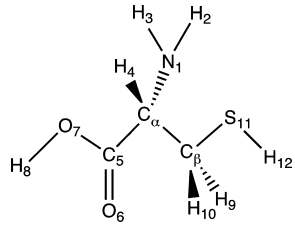


Figure 1. Labeling scheme for cysteine.

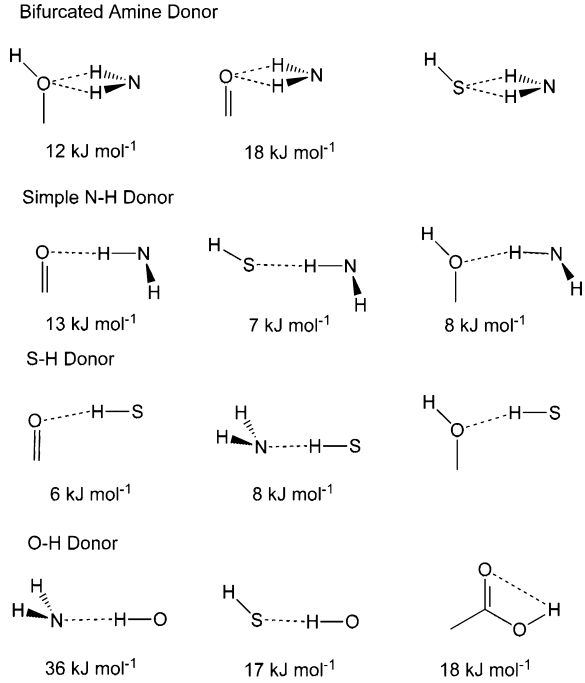


Figure 2. Common hydrogen bond motifs and approximate interaction strengths, where known.[15,22] Approximate relative energies of conformers can be treated additively as the difference between the sum of near-atom interactions.

structural and spectroscopic studies. The most accurate equilibrium structures (in cases both Born−Oppenheimer and semiexperimental ones) and relative energies (obtained within the focal-point analysis (FPA) approach[7−12]) are available for the amino acids glycine (Gly),[5,13,14] alanine (Ala),[13,15] threonine (Thr),[16] and proline (Pro).[17,18] The present high-level computational study expands the list of structurally well characterized amino acids by careful investigation of all of the low-energy conformers of L-cysteine.

Cysteine is a good representative of those amino acids with the added complexity of a side chain capable of substantial hydrogen-bonding interactions. Substitution of an -SH group for one of the H atoms of the methyl side chain of Ala introduces two new rotators of significance, those around the $C_\alpha-C_\beta$ and $C_\beta-S$ bonds (Figure 1). A dramatic increase in the number of possible conformers results. The presence of three H-bond donors and four H-bond acceptors in Cys allows for the existence of twelve distinct types of hydrogen bonds (Figure 2), including (a) bifurcated H-bonds between $-NH_2$ and −COOH, similar to those found in the most stable conformers of Gly[5] and Ala;[15] (b) simple N−H donor bonds, like the adducts to side chain S−H and to carboxylic acid O−H and C═O; (c) side chain S−H interactions with nitrogen and the carbonyl or hydroxyl oxygen; and (d) O−H donations to C═O and to side chain S−H and NH_2. While H-bonds are certainly the main secondary interactions determining the occurrence and relative energies of the conformers of Cys, other structural factors are also important. These include exchange, electrostatic, and hyperconjugative electronic effects, as well as steric and dispersive interactions. Detailed investigation of these structural factors is one of the main goals of this study. In principle, Cys could contain the same conformers as serine (Ser), its OH analogue. However, since the interactions in Cys are weaker than in Ser (the thiol group of the side chain has comparatively poorer H-bonding characteristics), the barriers separating the conformers are expected to be smaller and in some instances may even disappear. In such a case, Cys would exhibit fewer unique conformers than Ser.

Previous ab initio electronic structure computations performed on Cys include studies on its conformational behavior[19] and on its various physical properties, including proton affinities and ionization potentials.[20] Schäfer et al.[21] investigated 10 conformers of Cys at the RHF/4−21G level and established conformational trends. Gronert and O'Hair[22] located 42 conformers at several levels of ab initio electronic structure theory, including RHF/6-31G* and MP2/6-31+G*.[23] Recently, Dobrowolski and co-workers[24] located 51 conformers using the B3LYP and MP2 methods in conjunction with the aug-cc-pVDZ (and in some cases aug-cc-pVTZ) basis set. The computed B3LYP/aug-cc-pVDZ frequencies were compared to IR matrix isolation spectra, suggesting the presence of between three and six L-cysteine conformers in the experiments.

The conformers of cysteine investigated previously range in relative energy by at most 50 kJ mol^{-1}. The six most stable conformers of Cys lie within 7 kJ mol^{-1}, while altogether 33 conformers have been identified within a 17 kJ mol^{-1} range. Alonso et al.[25] recently identified five conformers within 10 kJ mol^{-1} using laser ablation and Fourier-transform microwave spectroscopy (FTMW). It is clear that the highest possible levels of electronic structure theory must be employed to obtain definitive energetics for these structurally diverse but energetically similar conformers.

The present study yielded, as primary information, accurate equilibrium structures and relative energies, as well as copious spectroscopic molecular parameters related to the vibrational and rotational spectra of the most important conformers of Cys. In turn, the large number of computed molecular properties allowed the investigation of a number of interesting computational issues. These include systematic errors in the geometries, bracketing the errors in relative energies for different levels of theory, anharmonicity and zero-point vibrational corrections, and the electron correlation effects in properties such as quadrupole coupling constants.

2. Computational Details

Most of the atom-centered Gaussian basis sets selected for the electronic structure computations of this study contain both polarization and diffuse functions, both of which are needed for the determination of accurate structures and relative energies of H-bonded systems.[26] The subcompact 3-21G[27,28] basis lacks these functions, and thus it was used

only for prescreening the conformers at the Hartree−Fock[29] level of theory. The correlation-consistent, polarized-valence (aug)-cc-p(C)V(n+d)Z, n = 2 (D), 3 (T), 4 (Q) basis sets of Dunning and co-workers[30−34] were employed extensively for geometry optimizations and single-point energy computations within the FPA approach.[7−10] The augmented (aug) basis sets contain diffuse functions, while tight functions necessary for treating core correlation are contained in the core-polarized (C) basis sets. In addition, the "+d" notation indicates a set of tight d-functions for second-row atoms to rectify problems with the originally designed correlation-consistent sets and thus smooth basis set extrapolations for sulfur-containing molecules. For Cys, the aug-cc-pV(D+d)Z, aug-cc-pV(T+d)Z, aug-cc-pV(Q+d)Z, and aug-cc-pV(5+d)Z basis sets contain 233, 492, 891, and 1458 CGFs, respectively. Only pure spherical harmonics were employed in all basis sets used in this study.

Electronic wave functions were determined in this study by the single-configuration, self-consistent-field, restricted Hartree−Fock (RHF) method,[29,35,36] by second-order Møller−Plesset perturbation theory (MP2),[23] and by coupled cluster (CC) methods,[37,38] including all single and double excitations (CCSD),[39] as well a perturbative correction for connected triple excitations [CCSD(T)].[40] In addition, energies and geometries were determined using the hybrid density functional B3LYP.[41−43] Both the T_1 diagnostics of coupled cluster theory[44,45] (∼0.014) and qualitative bonding principles indicate that the conformers of Cys are well described by single-reference correlation methods. The seven lowest 1s-like orbitals along with the sulfur 2s and 2p orbitals were kept as frozen core (FC) in all post-Hartree−Fock treatments.

The electronic structure packages MAB-ACESII,[46] MPQC,[47−49] MOLPRO,[50] and Gaussian03[51] were used extensively in this study.

2.1. Geometry Optimizations. Initial structures for the geometry optimizations of the conformers of Cys were found by systematically varying the six most important dihedral angles (see Figure 1). The thiol carbon, $\tau(S_{11}-C_\beta-C_\alpha-C_5)$, and amine group, $\tau(H_3-N_1-C_\alpha-C_\beta)$, were rotated in 30° increments, while the carboxylic acid plane, $\tau(O_6-O_7-C_5-C_\alpha)$, thiol hydrogen, $\tau(H_{12}-S_{11}-C_\beta-C_\alpha)$, carboxyl hydrogen, $\tau(H_8-O_7-C_5-O_6)$, and $C_\alpha-C_\beta$ bond, $\tau(S_{11}-C_\beta-C_\alpha-N_1)$, were rotated in 120° increments, resulting in a preliminary set of 11 664 starting structures. The initial geometries were optimized at the HF/3-21G level until the Cartesian displacements between optimization steps were less than 10^{-4} bohr. Redundant conformers were identified by checking that energies and geometries were equivalent within a given threshold. Energies were considered to be the same if they were within 10^{-7} E_h, while bond lengths and angles were required to be within 0.001 Å and 1.0°, respectively. In total, a set of 90 unique HF/3-21G conformers were found.

The HF/3-21G structures were reoptimized at the frozen-core MP2/cc-pVTZ level. When MP2(FC)/cc-pVTZ geometry optimizations were performed, some of the higher-energy HF/3-21G conformers disappeared, yielding a final set of 71 conformers for Cys, according to the same uniqueness criteria given above. The eleven MP2/cc-pVTZ conformers within 10.0 kJ mol^{-1} of the lowest-energy

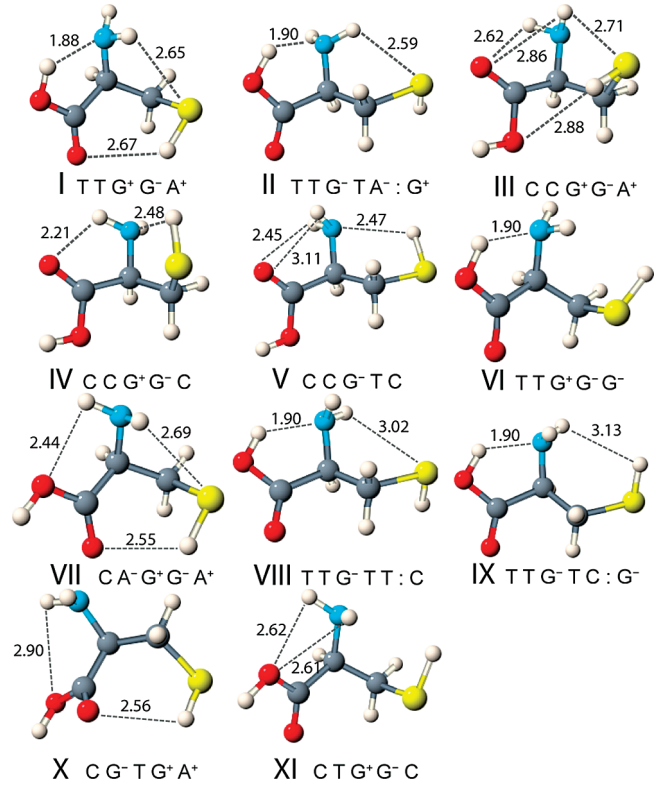


Figure 3. Pictorial representation of the eleven lowest-energy conformers of L-cysteine (Cys). See Figure 1 for numbering. The labeling scheme identifies conformers based on the series of dihedral angles $\tau(H_8-O_7-C_5-O_6)$, $\tau(N_1-C_\alpha-C_5-O_6)$, $\tau(N_1-C_\alpha-C_\beta-S_{11})$, $\tau(C_5-C_\alpha-C_\beta-S_{11})$, and $\tau(N_1-C_\alpha-S_{11}-H_6)$. A sixth identifier can be added for the "dihedral angle" of the nitrogen lone pair relative to the C_α-C_5 bond in cases of ambiguity (e.g., **II** and **VIII**).

structure were chosen for a more detailed analysis. These structures were reoptimized at the frozen-core MP2/aug-cc-pV(T+d)Z level. Detailed information, including Cartesian coordinates and energies of all the conformers found in this study, is provided as Supporting Information. Following a scheme first employed for glycine,[5] the conformers are numbered by Roman numerals (see Figure 3), reflecting the energy ordering determined at the MP2(FC)/aug-cc-pV(T+d)Z level. Similarly to Dobrowolski et al.,[24] we choose a series of dihedral angles to uniquely identify conformers, assigning each angle as C (cis, $-30° < \tau < +30°$), T (trans, $150° < \tau < 210°$), G$^+$ (gauche, $+30° < \tau < +90°$), G$^-$ (gauche, $-90° < \tau < -30°$), A$^+$ (anticlinal, $+90° < \tau < +150°$), or A$^-$ (anticlinal, $-150° < \tau < -90°$). This notation is equivalent to the Klyne−Prelog specification,[52] but we use the terms cis, gauche, and trans instead of synperiplanar, synclinal, and antiperiplanar, respectively. In particular, we identify the conformers via the dihedral angles $\tau(H_8-O_7-C_5-O_6)$, $\tau(N_1-C_\alpha-C_5-O_6)$, $\tau(S_{11}-C_\beta-C_\alpha-N_1)$, $\tau(S_{11}-C_\beta-C_\alpha-C_5)$, and $\tau(N_1-C_\alpha-S_{11}-H_{12})$. $\tau(H_8-O_7-C_5-O_6)$ specifies the carboxyl group as cis or trans, $\tau(N_1-C_\alpha-C_5-O_6)$ identifies the type of carboxyl-amine hydrogen bond, $\tau(S_{11}-C_\beta-C_\alpha-N_1)$ and $\tau(S_{11}-C_\beta-C_\alpha-C_5)$ define the backbone of the molecule, and $\tau(N_1-C_\alpha-S_{11}-H_{12})$ gives the orientation of the thiol hydrogen relative to the amine. For some conformers, for example, **Cys-II** and **Cys-VIII**, the

orientation of the amine lone pair relative to the C_5-C_α bond must also be specified to uniquely identify the conformer.

The 71 conformers located here exceed those of Dobrowolski et al.,[24] and Gronert and O'Hair,[22] who observed only 51 and 42 distinct Cys conformers, respectively. Their initial search tested 324 starting geometries at the AM1 level, resulting in a preliminary set of only 58 conformers. In contrast, our initial search was performed at the HF/3-21G level on an initial set of over 10 000 starting geometries, yielding a preliminary set of 90 conformers. Because it is highly likely that the 71 distinct MP2(FC)/cc-pVTZ conformers located here exist in reality, it seems that the 20 conformers were missed in previous work due to the AM1 preoptimization step.

2.2. Focal-Point Analysis (FPA). To obtain relative energies as accurately as possible, the focal point analysis (FPA) approach[7−12] was utilized. The eleven lowest-energy structures of Cys (Figure 3), obtained at the MP2(FC)/aug-cc-pV(T+d)Z level, were included in the FPA investigation. Extrapolation of the energies to the complete basis set (CBS) limit at the RHF and MP2 levels was performed, as part of the FPA approach. For HF, the total energy was extrapolated using the formula $E_n = E_{CBS} + A\exp(-Bn)$[53−55] with $n \in \{3,4,5\}$, where A and B are adjustable parameters, E_n is the RHF total energy for a correlation-consistent basis set aug-cc-pV(n+d)Z, and E_{CBS} is the Hartree−Fock limit. For MP2, the correlation energies (ε_n) were extrapolated according to $\varepsilon_n = \varepsilon_{CBS} + Bn^{-3}$.[56] Coupled-cluster energy corrections were treated additively because the electron correlation contributions to the conformational energies at higher levels of theory do not change significantly with the size of the basis set.

For the auxiliary corrections normally included within the FPA approach, the core correlation term was obtained at the MP2/aug-cc-pCVTZ level, while the relativistic[57] and diagonal Born−Oppenheimer corrections (DBOC)[58] were deemed negligible. In similar FPA studies performed on the conformers of proline[18] and threonine,[16] the relativistic corrections to the relative conformational energies turned out to be minuscule.

2.3. Spectroscopic Parameters. For the eleven lowest-energy conformers of Cys, quadratic force constants were computed at the B3LYP/aug-cc-pVTZ level using geometries optimized at the same level. In addition, quartic force fields in the normal coordinate space were determined at the B3LYP/6-31G* level. No scaling of the force fields or of the resulting vibrational frequencies was attempted. The use of fully optimized reference geometries during the force field determinations helps to avoid the nonzero force dilemma.[59] Anharmonic, fundamental vibrational frequencies of the Cys conformers were computed by applying the VPT2 formalism[60−63] to the quartic force fields. Whenever a Fermi resonance appeared, the corresponding contribution to the fundamental frequency was evaluated by eliminating the associated terms in the expression for the anharmonic constants and then explicitly diagonalizing the 2 × 2 Hamiltonian matrix for the resonating states.

The optimized structures obtained at the MP2(FC)/aug-cc-pVTZ level determine the equilibrium rotational constants, while the quadratic and cubic force fields from B3LYP/6-31G* yield the quartic and sextic centrifugal distortion constants in the A-reduced representation, respectively. In additition, quadrupole coupling constants are reported at the MP2 level using a "locally dense" basis set composed of the standard cc-pVTZ functions on carbon, sulfur, oxygen, and hydrogen and the cc-pCV5Z functions on nitrogen. The quadrupole coupling constant is determined by the electric field gradient at nitrogen so that the combination of a locally dense basis and MP2 generally produces good results (see Supporting Information).

2.4. Zero-Point Vibrational Corrections. Zero-point vibrational energies (ZPVE) were obtained as harmonic and anharmonic values at the B3LYP/aug-cc-pVTZ and B3LYP/6-31G* levels, respectively. Some ambiguity exists in the computation of zero-point energies based on anharmonic force fields. For molecules with many normal modes, resonances occur in the energy denominators of the related second-order vibrational perturbation theory (VPT2)[60−63] expressions. Here, we employ the approach of Allen et al.,[64] in which the oft-neglected G_0 term is included to obtain an expression for the zero-point vibrational energy completely devoid of resonance denominators. The working equations are

$$\mathrm{ZPVE} = \frac{1}{2}\sum_i \omega_i - \frac{1}{32}\sum_{ijk}\frac{\phi_{iik}\phi_{jjk}}{\omega_k} - \frac{1}{48}\sum_{ijk}\frac{\phi_{ijk}^2}{\omega_i+\omega_j+\omega_k} + \frac{1}{32}\sum_{iijj}\phi_{iijj} + Z_{\mathrm{Kinetic}} \quad (1)$$

and

$$Z_{\mathrm{Kinetic}} = -\frac{1}{4}\sum_\alpha B_e^\alpha\left\{1 + \sum_{i>j}(\zeta_{ij}^\alpha)^2\frac{B_e^\alpha(\omega_i+\omega_j) - (\omega_i-\omega_j)^2}{\omega_i\omega_j}\right\} \quad (2)$$

where the ϕ_{ijk} and ϕ_{ijkl} are cubic and quartic force constants in reduced normal coordinates, B_e^α denotes the rotational constant for axis α, and the ζ_{ij}^α are Coriolis coupling constants. The total zero-point vibrational energies obtained directly from Gaussian 03 were quite similar to those from eq 1, but contained discrepancies as large as 30 cm^{-1} for some conformers. While both our method and Gaussian 03 include the G_0 term, Gaussian 03 includes anharmonic effects by summing over the average of the harmonic and fundamental frequencies[65]

$$\mathrm{ZPVE} = \frac{1}{4}\sum_i(\omega_i + \nu_i) + G_0 - \frac{1}{4}\sum_i x_{ii} \quad (3)$$

where x_{ii} is the diagonal anharmonic constant for mode i. The anharmonic corrections to the ZPVE in Gaussian 03 therefore avoid resonance denominators by including the explicit 2 × 2 matrix diagonalization for resonating states (section 2.3), in contrast to our approach which avoids resonance denominators by computing ZPVE directly in terms of cubic and quartic force constants.

Anomalously large VPT2 anharmonic shifts are observed for rotation of the S−H bond in the conformers **Cys-IV**, **Cys-VI**, and **Cys-XI**. In general, for conformers **Cys-IV**, **Cys-VI**, and **Cys-XI**, the thiol hydrogen is near the amino group,

Table 1. Summary of Focal Point Analysis for the Relative Energies of the Eleven Most Stable Conformers of Cysteine[a]

	Basis	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
ΔE(RHF)[a]	DZ	0.00	3.82	−3.91	−0.09	−5.29	7.62	−2.85	3.23	4.57	−2.75	−0.17
	TZ	0.00	3.67	−4.00	−0.39	−5.40	7.22	−2.34	3.15	4.27	−2.52	−0.13
	QZ	0.00	3.59	−4.02	−0.50	−5.54	7.13	−2.33	2.97	4.06	−2.54	−0.17
	5Z	0.00	3.57	−4.04	−0.54	−5.59	7.11	−2.36	2.94	4.02	−2.57	−0.23
	CBS	0.00	3.56	−4.06	−0.55	−5.61	7.09	−2.39	2.93	4.01	−2.60	−0.25
δ[MP2][a]	DZ	+0.00	+2.20	+10.67	+8.37	+12.77	+1.80	+11.14	+7.24	+6.90	+12.64	+11.25
	TZ	+0.00	+2.83	+11.43	+8.98	+14.38	+2.07	+11.64	+7.87	+7.50	+14.46	+12.32
	QZ	+0.00	+2.94	+11.71	+9.00	+14.40	+2.28	+11.72	+7.82	+7.49	+14.60	+12.62
	5Z	+0.00	+2.98	+11.74	+8.98	+14.41	+2.32	+11.69	+7.84	+7.52	+14.65	+12.63
	CBS	+0.00	+3.02	+11.78	+8.95	+14.43	+2.37	+11.66	+7.86	+7.55	+14.71	+12.64
δ[CCSD]	DZ	+0.00	−0.66	−2.90	−1.84	−3.20	−1.03	−2.92	−2.60	−2.56	−3.32	−3.04
	TZ	+0.00	−0.70	−3.05	−2.03	−3.64	−1.00	−3.14	−2.83	−2.76	−3.60	−3.30
δ[CCSD(T)]	DZ	+0.00	+0.42	+1.88	+1.64	+2.04	+0.30	+1.80	+1.08	+1.01	+1.84	+1.85
	TZ	+0.00	+0.49	+2.06	+1.78	+2.33	+0.32	+1.96	+1.15	+1.07	+2.21	+2.10
ΔE(CCSD(T)/CBS)		0.00	6.37	6.73	8.15	7.51	8.78	8.09	9.11	9.87	10.72	11.19
core correction		+0.00	+0.13	+0.17	+0.04	+0.33	+0.04	+0.17	+0.21	+0.21	+0.33	+0.21
harmonic ZPVE		+0.00	−0.01	−2.19	−2.38	−2.24	−0.71	−1.73	−0.93	−1.00	−1.85	−1.76
anharmonic correction		+0.00	+0.08	+0.08	+0.14	+0.21	−0.55	−0.01	+0.21	+0.05	+0.32	−0.10
ΔE(FPA)		0.00	6.58	4.79	5.95	5.81	7.56	6.52	8.60	9.13	9.52	9.54

[a] All values given in kJ mol^{-1}. ΔE denotes a relative energy between conformers. δ denotes an incremement or correction to ΔE with respect to the preceding level of theory in the hierarachy RHF → MP2 → CCSD → CCSD(T).

and the thiol rotation is strongly coupled to the NH_2 wag. Because of this coupling, VPT2 breaks down for these modes, resulting in unphysically large anharmonic shifts. In eq 1, we exclude those force constants involving at least one mode for which VPT2 gives an anomalous fundamental frequency.

2.5. Natural Bond Orbital Analysis. The natural bond orbital (NBO) method transforms the molecular orbital picture into a localized description based on the intuitive Lewis structures of molecules.[66−70] The NBO scheme decomposes the 1-particle density matrix into formally occupied 1-center core orbitals and lone pairs (n_x) and 2-center bonding orbitals (σ_{X-Y}, π_{X-Y}), which correspond naturally to bonds drawn in a Lewis diagram. The localized, Lewis density is transformed to the exact density through delocalizations into formally unoccupied 2-center antibonding orbitals (σ^*_{X-Y}, π^*_{X-Y}) and unoccupied 1-center Rydberg orbitals. These delocalizations can be interpreted physically as energy-stabilizing donor−acceptor interactions between localized orbitals, including conjugation (e.g., $\pi \rightarrow \pi^*$) or hyperconjugation ($\sigma \rightarrow \sigma^*$). Applying perturbation theory gives the leading, second-order energy correction due to these donor−acceptor interactions as

$$E_2 = -\frac{F_{ij}^2}{\varepsilon_i - \varepsilon_{j^*}} \tag{4}$$

where F_{ij} is the Fock matrix element between occupied orbital i and unoccupied orbital j^*, and ε_i and ε_{j^*} are the corresponding diagonal Fock matrix elements. Here i and j^* do not represent canonical doubly occupied and virtual molecular orbitals, but rather "almost doubly occupied" and "almost unoccupied" orbitals, respectively. In this regard, the Brillouin condition does not apply so that the matrix elements F_{ij}, while small, are not rigorously zero. Written this way, the E_2 values indicate the most important relaxation from an idealized, local density to the exact density. For cysteine, we can therefore assess the importance of hyperconjugation in conformational transitions through the NBO formalism. Because E_2 is based on perturbation theory, it may overestimate the absolute magnitude of strong interactions in which the energy denominator is small or the Fock matrix element is large. However, the overall trends should be consistent between conformers.

3. Results and Discussion

3.1. Relative Energies of Conformers. As observed repeatedly for amino acids, the introductory Hartree−Fock level of electronic structure theory, independent of the basis set used, is unable to yield the correct relative energies of the conformers of Cys (Table 1). RHF/CBS theory places six conformers (**Cys-III**, **-IV**, **-V**, **-VII**, **-X**, and **-XI**) lower in energy than the global minimum, **Cys-I**. While in Gly and Ala the inclusion of electron correlation tends to decrease the energy differences between the conformers,[5,15] in the case of Cys it increases the energy differences in almost all cases, which might be attributed to the relatively weak interactions present in Cys. Because sulfur is much more polarizable than oxygen or carbon, the dipole−induced-dipole and dispersion forces should be generally more important in Cys than in Ala, Ser, or Pro, while S−H hydrogen bonds should be weaker. Accordingly, the MP2 correlation energy destabilizes all the conformers considered relative to **I,** as signified by the positive δ[MP2] values in Table 1, which can be as large as 15 kJ mol^{-1}. This observation also serves as a warning that the theoretical results obtained with small basis sets and simple electronic structure methods might change considerably once more rigorous techniques are employed.

Compared to δ[MP2], the δ[CCSD] and δ[CCSD(T)] energy increments are relatively small but can affect relative energies as much as 3.6 and 2.3 kJ mol^{-1} for CCSD and CCSD(T), respectively. Such amounts are clearly substantial when so many conformers are within a window of a few kJ mol^{-1}. Interestingly, but again in line with earlier work on the conformers of Thr[16] and Pro,[18] the relative energies are barely affected by core correlation. Even the largest change is smaller than 0.35 kJ mol^{-1}. In contrast, ZPVE corrections

Table 2. Comparison of Conformational Energies of Cysteine (kJ mol^{-1}) for Different Levels of Theory without Zero-Point Vibrational Correction

	CCSD(T)/CBS[a,b]	MP2/aug-cc-pV(T+d)Z[b]	CCSD/aug-cc-pV(T+d)Z[b]	RHF/3-21G[c]	B3LYP/aug-cc-pVTZ[d]
Cys-I	0.00	0.00	0.00	1.26	0.00
Cys-II	6.37	6.49	5.82	11.84	3.93
Cys-III	6.73	7.45	4.35	4.35	7.15
Cys-IV	8.15	8.58	6.57	2.43	8.79
Cys-V	7.51	9.00	5.36	0.00	5.44
Cys-VI	8.78	9.29	8.28	12.68	7.53
Cys-VII	8.09	9.29	6.15	5.23	6.36
Cys-VIII	9.11	11.00	8.16	9.08	6.36
Cys-IX	9.87	11.76	9.00	11.30	7.15
Cys-X	10.72	11.97	8.37	13.05	8.91
Cys-XI	11.19	12.18	8.87	10.04	11.09

[a] CCSD(T)/CBS denotes the extrapolated value from the focal point analysis. See Table 1. [b] Computed at the MP2/aug-cc-pV(T+d)Z reference geometries. [c] Computed at the RHF/3-21G reference geometries. [d] Computed at the B3LYP/aug-cc-pVTZ reference geometries.

can affect relative energies on the order of 2 kJ mol^{-1}. Anharmonic corrections to the relative energies were less than 0.21 kJ mol^{-1} for all conformers except **Cys-VI** and **Cys-X**, for which these shifts were −0.55 and +0.32 kJ mol^{-1}, respectively.

The MP2(FC)/aug-cc-pV(T+d)Z single-point energies happen to be quite accurate (Table 2) because the CCSD and CCSD(T) increments are usually of opposite sign and partially cancel (Table 1). In this regard, many of the MP2 relative energies in Table 2 are closer to the FPA results than their CCSD counterparts. The effect of higher-order correlation from the CCSD(T) perturbative triples is not negligible, however, and the definitive FPA scheme alters the MP2 energy ordering of the conformers. In general, B3LYP performs reasonably well for most conformers but can be in error by as much as 2.7 kJ mol^{-1}, as seen in **Cys-VIII**. While density functional theory can be useful for zero-point corrections and geometry optimizations, obtaining the correct energy ordering of so many conformers in such a small energy range (10 kJ mol^{-1}) clearly requires better accuracy than B3LYP provides. A rigorous energy ordering therefore necessitates correlation treatments as extensive as CCSD(T) and also considerations inherent in the FPA scheme.

The FPA scheme allows errors to be bracketed based on the observed convergence to the basis set and correlation limits. The RHF relative energies are converged to better than 0.05 kJ mol^{-1} with the aug-cc-pV(5+d)Z basis, and thus there should be virtually no basis set error in our final RHF/CBS results. Similarly, the MP2/aug-cc-pV(5+d)Z correlation increments match the extrapolated values within 0.1 kJ mol^{-1} for all conformers, and the associated basis set errors should again be negligible. For the coupled-cluster increments, the aug-cc-pV(T+d)Z result matches the aug-cc-pV(D+d)Z result within 0.4 kJ mol^{-1}. In previous work, the CCSD and CCSD(T) increments are essentially converged with a TZ basis,[71] so that the basis set error in the coupled-cluster values should not be greater than 0.2 kJ mol^{-1}.

Assesssing the error caused by the higher-order correlation and zero-point vibrational corrections is more difficult. Corrections due to quadruple and higher excitations are generally an order of magnitude less than CCSD(T) corrections.[71−78] Since most CCSD(T) corrections here are on the order of 1−2 kJ mol^{-1}, the neglect of higher excitations should introduce an error to the cysteine relative energies of at most 0.2 kJ mol^{-1}. Zero-point vibrational corrections are generally insensitive to the level of theory. For example (see Supporting Information, Table S1), even MP2 harmonic zero-point corrections with the Huzinaga−Dunning DZP++ basis[79] (double-ζ plus polarization and diffuse functions) match the B3LYP/aug-cc-pVTZ values within 0.3 kJ mol^{-1}. Since we have accounted for anharmonicity in the present work, the zero-point error should therefore not be greater than 0.2−0.3 kJ mol^{-1}. In summary, we estimate a standard error (1σ) of 0.3 kJ mol^{-1} for our predicted conformational energies, corresponding to a 95% confidence interval (2σ) of ± 0.6 kJ mol^{-1}. We emphasize that these uncertainties hold only for the relative energies due to cancelation of errors, and the uncertainty in the absolute energies will therefore be much larger.

Two recent studies published relative energies for the lowest-energy conformers of cysteine as summarized in the Supporting Information. All eight conformers considered by Dobrowolski et al.[24] were also studied in the current work. Three conformers from Alonso et al.[25] were not within 10 kJ mol^{-1} of the global minimum after optimization at the MP2(FC)/cc-pVTZ level, and were therefore not included in our rigorous focal point analyses. All three studies agree in the structure of the two most stable conformers of cysteine, **Cys-I** and **Cys-III** in our notation. The ordering of the other low-lying conformers is similar but not the same in the three studies, and the relative energies vary by as much as 1.3 kJ mol^{-1} for **Cys-III** and 1.9 kJ mol^{-1} for **Cys-IX**.

3.2. Geometric Structures. The sophisticated laser ablation FTMW experiments of Alonso et al.[25] yielded rotational constants of several conformers of cysteine, but only for the parent isotopologues. Therefore, the type of refinement on collections of isotopologues which has yielded semiexperimental equilibrium structures for Gly[14] and Pro[17] cannot be executed at present for any of the conformers of Cys. Consequently, one must rely on otherwise rather accurate[15,17] computed structures when analyzing structural trends among the conformers of Cys.

Two major factors seem to determine the general type of conformation that Cys can assume. First, the thiol, amine, and hydroxyl groups adopt different orientations about the $C_\alpha-C_\beta$ bond as either gauche or trans. In the discussion to

follow, unless stated otherwise, gauche and trans identify the orientation about the C_α-C_β bond. Second, the carboxyl group may assume a cis or trans conformation. Depending on these orientations, different hydrogen bonding patterns can form of the types O−H···N, N−H···O═C, and N−H···OH, as clearly seen in Figure 3. For the trans carboxyl, a strong O−H···N interaction can form as found in conformers **Cys-I**, **Cys-II**, **Cys-VI**, **Cys-VIII**, and **Cys-IX**. For the cis carboxyl, bifurcated N−H···O bonds form to either the carbonyl oxygen in **Cys-III**, **Cys-IV**, and **Cys-V** or to the hydroxyl oxygen in **Cys-VII** and **Cys-XI**. The gauche conformers are more sterically crowded than the trans conformers. However, the gauche conformation brings the thiol group closer to the carboxyl group, allowing S−H···O interactions. A ring of hydrogen bonds can therefore form, as in **Cys-I**, **Cys-III**, and **Cys-VII**. For trans conformers, only the amine interacts with the thiol, for example in **Cys-II**.

As observed previously for other amino acids,[5,15,18] bond lengths and bond angles change little among the low-lying conformers. Most bonds have a standard deviation of less than 0.004 Å, while most bond angles have a standard deviation of less than 2.0°. There are, however, a few notable exceptions. The C═O distance has a standard deviation of 0.007 Å with the largest deviation from the mean being 0.012 Å, occurring in **Cys-III**. The C═O bond in **Cys-III** forms a bifurcated hydrogen bond with the amine group, lengthening the bond and redshifting the carbonyl stretching frequency. The C_α−C_5 bond has a standard deviation of 0.008 Å, with the largest deviation from the mean (0.012 Å) occurring in **Cys-IV**. In general, three strong hyperconjugative interactions are possible for the C_α−C_5 bond with the lone pairs from nitrogen, the hydroxyl oxygen, and the carbonyl oxygen. The hyperconjugation is strongest when the lone pair is trans to the C_α−C_5 bond (see below), and it will lengthen this bond by increasing the antibonding occupation. In **Cys-IV**, both the amine and hydroxyl groups are unfavorably placed for hyperconjugation, and the C_α−C_5 bond distance is only 1.514 Å. In contrast, both the hydroxyl and amine are favorably placed for hyperconjugation in **Cys-I**, lengthening the C_α−C_5 bond to 1.534 Å. Similarly, in **Cys-III**, the amine lone pair is trans to the C_α−C_5 bond, but the hydroxyl lone pair is unfavorably placed cis. The C_α−C_5 bond therefore has an intermediate value of 1.521 Å.

For bond angles, the largest standard deviations are for the N−C−C angles. The standard deviation for N_1−C_α−C_β is 2.8° with the largest deviation from the mean being 4.0° for **Cys-VI**. Similarly, the standard deviation for N_1−C_α−C_5 is 2.8° with the largest deviation from the mean value being 5.4° for **Cys-XI**. The large spread of N−C−C angles is consistent with the trans angle rule.[80] In general, in primary alcohols and amines, if a C−C bond is trans to the X−H bond, the X−C−C angle will be smaller, because of both reduced bond repulsion relative to the gauche conformer and weaker hyperconjugation from the nitrogen lone pair. The change in angle depending on bond orientation is clearly evident in cysteine. For conformers **Cys-I**, **Cys-VI**, **Cys-VIII**, and **Cys-IX**, the N−H bonds are gauche to the C_α-C_β bond, and the angles are in the range 115°−117° (see Supporting Information Table S3). In contrast, for conformers **Cys-II**, **-III, -IV**, **-V**, **-VII**, **-X**, and **-XI**, the N−H bond is trans to the C_α−C_β bond, and the N−C−C angles range from 109° to 111°. The same general trends are observed for the N_1−C_α−C_5 angles. The angle variations are consistent with strong hyperconjugation from the nitrogen lone pair to the C−C antibonding orbital ($n_N \rightarrow \sigma^*_{C-C}$). Following arguments rationalizing tilting of the methyl group,[81,82] the C−C axis generally tilts away from the C−N axis to maximize overlap between the nitrogen lone pair and the backside lobe of the C−C antibonding orbital, strengthening the $n_N \rightarrow \sigma^*_{C-C}$ hyperconjugative stabilization. In particular, for both C_α−C_5 and C_α−C_β, large hyperconjugative interactions ($E_2 > 20.0$ kJ mol^{-1}, Supporting Information Table S3) are observed with large C−C−N angles, while weaker interactions ($E_2 <$ 16.0 kJ mol^{-1}) are observed with smaller C−C−N angles.

In the same way for the carboxyl group, if the carboxyl group assumes a cis conformation (O−H bond trans to the C−H bond), the O−C−C angle is much smaller. This is observed in conformers **Cys-III**, **Cys-IV**, **Cys-V**, and **Cys-VII**, all of which have O−C−C angles of approximately 111.5°. In contrast, for conformers **Cys-I**, **Cys-II**, **Cys-VI**, and **Cys-VIII** with trans carboxyl (O−H bond cis to the C−C bond), the angles are larger, between 113.5° and 114°. In general, the trans effect seems weaker for the O−H bond than for the N−H bond. The weaker dependence may be attributed to the stronger basicity of the amine and therefore stronger $n \rightarrow \sigma^*$ hyperconjugation. The O−H···N hydrogen bonding also seems to offset the trans effect, closing the O−C−C angle to maximize the O−H···N interaction.

3.3. Structural Effects on Relative Energies. As emphasized previously for alanine[15] and serine,[22] approximate values for the strength of certain types of hydrogen bonds can be computed and used to rationalize energy differences among the conformers. Specifically, the energy of each conformer can be approximated as a sum of stabilizations from near-atom interactions. The interaction strengths are then fit through a linear regression to match as closely as possible the conformational energies. Approximate hydrogen bond strengths are given in Figure 2 for the common bonding motifs, as reported in ref 15. In general, the hydrogen bond donors can be ranked in the order O−H > N−H > S−H, and the hydrogen bond acceptors can be ranked in the order N > O > S. Additionally, an additive approximation can be applied to the conformation of the carboxyl group. Based on the formic acid prototype,[15] the cis carboxyl is intrinsically more stable than the trans carboxyl by approximately 18.5 kJ mol^{-1} irrespective of hydrogen bonds to other functional groups.

As shown in Figure 2, the O−H···N arrangement is the strongest hydrogen bond, matching the strongest hydrogen bond donor, OH, with the best acceptor, N. We therefore find that the structure of the global minimum, **Cys-I**, is stabilized by a strong O−H···N hydrogen bond between the amino and carboxyl groups. The Cys global minimum is in contrast to that of serine,[22] for which the lowest energy conformer is analogous to **Cys-V**, exhibiting a strong O−H···N hydrogen bond with the side chain. The S−H bond in Cys is comparatively a much weaker hydrogen bond

Table 3. Anharmonic Vibrational Fundamentals in cm^{-1} and Double-Harmonic Infrared Relative Intensities (%) of the Eleven Lowest-Energy Conformers of L-Cysteine for Regions Characteristic of Particular Hydrogen Bond Patterns[a]

	O−H, N−H stretch			S-H stretch	C=O stretch	O−H bend		C−O stretch	
	ν_1	ν_2	ν_3	ν_7	ν_8	ν_{11}	ν_{12}	ν_{17}	ν_{18}
Cys-I	3397 (4)	3278 (18)	3216 (47)	2542 (0)	1793 (75)	1356 (100)	1353 (2)	1129 (2)	1073 (3)
Cys-II	3383 (6)	3300 (15)	3226 (43)	2543 (0)	1807 (87)	1337 (100)	1337 (8)	1144 (3)	1068 (3)
Cys-III	3539 (22)	3387 (4)	3327 (1)	2551 (0)	1776 (100)	1350 (0)	1314 (3)	1121 (2)	1086 (78)
Cys-IV	3554 (30)	3417 (8)	3374 (5)	2557 (1)	1775 (100)	1385 (4)	1314 (9)	1114 (69)	1084 (22)
Cys-V	3541 (21)	3397 (3)	3324 (1)	2552 (1)	1773 (100)	1380 (2)	1278 (15)	1124 (15)	1084 (58)
Cys-VI	3397 (4)	3378 (54)	3238 (12)	2539 (0)	1798 (84)	1349 (100)	1353 (12)	1130 (4)	1065 (2)
Cys-VII	3545 (28)	3407 (5)	3392 (1)	2548 (0)	1762 (100)	1350 (0)	1330 (8)	1114 (69)	1102 (22)
Cys-VIII	3396 (3)	3305 (10)	3244 (62)	2548 (0)	1794 (100)	1336 (59)	1366 (82)	1120 (3)	1084 (2)
Cys-IX	3410 (2)	3395 (18)	3254 (32)	2525 (0)	1798 (71)	1332 (1)	1362 (100)	1112 (5)	1088 (1)
Cys-X	3538 (25)	3398 (2)	3301 (0)	2560 (0)	1766 (100)	1344 (10)	1354 (0)	1111 (90)	1056 (21)
Cys-XI	3538 (27)	3401 (3)	3315 (0)	2545 (0)	1767 (100)	1359 (2)	1309 (10)	1122 (70)	1102 (18)

[a] Harmonic frequencies and intensities were computed at the B3LYP/aug-cc-pVTZ level. Anharmonic corrections were computed using B3LYP/6-31G*. Intensities are reported as a percentage of the most intense peak for a given conformer.

donor than the O−H side chain in serine so that the trans to cis isomerization of the carboxyl is not enough to offset the weaker hydrogen bond. In addition, **Cys-I**, because of its gauche conformation, can form three hydrogen bonds whose charge polarization will cooperatively reinforce each other. We emphasize that our focal point conformational energies should be accurate within a standard error of 0.3 kJ mol^{-1} (1σ) or a 95% confidence interval of ± 0.6 kJ mol^{-1} (2σ). In contrast to previous studies,[24,25] we can therefore definitively say that the energy differences are real physical effects rather than errors in the underlying computational methods.

In addition to hydrogen bonding, gauche and trans conformations are also affected by steric repulsion and hyperconjugation. The gauche conformers of cysteine have all three bulky substituents in close vicinity, increasing steric repulsion. For butane, the trans−gauche difference is 2.6 kJ mol^{-1}.[83] Consequently, steric effects are certainly not negligible in Cys since its eleven lowest-energy conformers lie in an energy range of 10 kJ mol^{-1}. Hyperconjugation is stronger in the gauche configuration since the strongly electronegative groups (amine, carboxyl, thiol) are all antiperiplanar to the electropositive hydrogens. In this regard, the better electron donor orbitals (σ_{C-H}) are matched to the better electron acceptors (σ^*_{C-N}, σ^*_{C-S}). The leading hyperconjugative interactions in cysteine conformers are listed in Table S4, Supporting Information. For example, the $\sigma_{C-H} \rightarrow \sigma^*_{C-S}$ interaction is 20.7 kJ mol^{-1} in the gauche **Cys-I**, while the equivalent $\sigma_{C-H} \rightarrow \sigma^*_{C-H}$ interaction is only 11.3 kJ mol^{-1} in **Cys-V,** although the effect is offset somewhat by stronger $\sigma_{C-C} \rightarrow \sigma^*_{C-S}$ hyperconjugation in **Cys-V** relative to the $\sigma_{C-C} \rightarrow \sigma^*_{C-H}$ hyperconjugation in **Cys-I.** This gauche effect has been observed previously for difluorosubstituted hydrocarbons and hydroxyproline.[84,85] Hyperconjugation and steric effects will therefore tend to offset each other. For some conformers, the relative energies will therefore be dictated mainly by the hydrogen bonding interactions, owing to fortuitious cancelation of competing electronic effects.

The importance of hyperconjugation can be seen in the transformation from **Cys-IV** to **Cys-V**, wherein the amine switches from a simple N−H···O bond to a bifurcated N−H···O bond, while the orientation about the $C_\alpha-C_\beta$ bond simultaneously goes from gauche to trans. Assuming simple hydrogen bond additivity, **Cys-V** should lie 5 kJ mol^{-1} below **Cys-IV** due to the larger bifurcated hydrogen bond energy (Figure 2). In fact, the energy difference is less than 1 kJ mol^{-1}. The S−H···N bond distances are basically equivalent in **Cys-IV** and **Cys-V** (2.48 Å versus 2.47 Å), so the S−H···N interaction should not contribute significantly to the energy difference. The discrepancy seems to lie in hyperconjugative stabilization of **Cys-IV**. While steric repulsion is greater in **Cys-IV**, the much stronger $\sigma_{C-H} \rightarrow \sigma^*_{C-S}$ and $\sigma_{C-H} \rightarrow \sigma^*_{C-C}$ donations in **Cys-IV** relative to the $\sigma_{C-C} \rightarrow \sigma^*_{C-S}$ and $\sigma_{C-S} \rightarrow \sigma^*_{C-C}$ delocalizations in **Cys-V** (Table S4, Supporting Information) preferentially stabilize **Cys-IV**.

Perhaps the most interesting conformer is **Cys-X**, which forms no typical hydrogen bonds in the sense of near-linear X−H···Y arrangements. The carboxyl plane is perpendicular to the C−N bond. In this way, the amine seems to form both weak N−H···O−H and N−H···O=C interactions. Despite positioning the C−N bond trans to the C−S bond, **Cys-X** exhibits strong $\sigma_{C-S} \rightarrow \sigma^*_{C-N}$ hyperconjugation, contributing to its unusual stability.

3.4. Vibrational Fundamentals. Anharmonic vibrational frequencies in characteristic hydrogen bonding regions of the infrared spectra are given in Table 3 for the eleven lowest-energy conformers of Cys. Complete sets of vibrational fundamentals for these conformers are provided in Table S5 of Supporting Information. The combination of B3LYP/aug-cc-pVTZ harmonic frequencies and B3LYP/6-31G* anharmonic corrections employed here should be accurate on average to within 30 cm^{-1}, although some deviations may be substantially larger.[86]

Various fundamentals can be found where either the intensities or the frequencies distinguish between structural features. For example, **Cys-I** is the only conformer that has all of the following: two medium/strong bands between 3200−3300 cm^{-1}, no strong bands above 3300 cm^{-1}, and only two medium/weak bands below 1300 cm^{-1}. In matrix isolation experiments on Cys, Dobrowolski et al.[87] noted that broadening and clustering of peaks often prevented assignment of absorptions to individual conformers. In general, the matrix isolation vibrational spectra were only able to distinguish between conformers as being with or without certain intramolecular hydrogen bonds. As a natural extension, Dobrowolski et al. suggested that vibrational circular dichroism (VCD) may be more informative. Because the

relatively accurate computation of VCD spectra is straightforward, such experiments hold promise for the detection of conformers present in the gas before matrix deposition.

Despite possible difficulties, four different regions should provide important fingerprints for cysteine conformers to distinguish between O−H···N or N−H···O═C bonding patterns. No conformers with strong N−H···O−H hydrogen bonds appeared in the current study. The first region is the O−H stretch region between 3200−3600 cm^{-1}. For conformers such as **Cys-I**, **Cys-II**, **Cys-VI**, and **Cys-VIII** with O−H···N hydrogen bonds, the O−H stretches are redshifted so that the highest frequency peaks are the N−H stretches near 3400 cm^{-1}. For free hydroxyl groups, the O−H stretch appears near 3540 cm^{-1}, as seen in conformers **Cys-III**, **Cys-IV**, **Cys-V**, **Cys-VII**, **Cys-X**, and **Cys-XI**. This general band structure was observed in the matrix IR study,[87] and our anharmonic fundamentals agree with the experimental absorptions within 10−20 cm^{-1}, consistent with the uncertainty estimate given above. As noted by Dobrowolski et al.,[87] the C−O single-bond stretch also provides an important diagnostic through differing intensitities. In conformers **Cys-I**, **Cys-II**, **Cys-VI**, and **Cys-VII** with O−H···N bonds, the C−O stretching region does not exhibit any high intensity peaks, presumably because the oscillator strength is smeared out among several modes. In contrast, for hydroxyl groups that do not form hydrogen bonds, the C−O stretch in **Cys-III**, **Cys-IV**, **Cys-V**, **Cys-VII**, **Cys-X**, and **Cys-XI** shows strong features between 1080 and 1125 cm^{-1}, again in excellent agreement with the experimentally observed frequencies.

As usual, the most telling band is the carbonyl stretch, ν_8, which ranges from 1762−1807 cm^{-1}. For conformers **Cys-III, Cys-IV, Cys-V, Cys-VII, Cys-X,** and **Cys-XI,** which contain a N−H···O═C hydrogen bond, the stretch is redshifted, appearing between 1762−1776 cm^{-1}. In contrast, the free carbonyls in **Cys-I**, **Cys-VI**, **Cys-VIII**, and **Cys-IX** all appear in a narrow range around 1795 cm^{-1}. The highest frequency occurs for **Cys-II** at 1807 cm^{-1}. However, no peak appears above 1800 cm^{-1} in the experimental spectrum.[87] **Cys-II** is essentially identical to **Cys-VIII** except for a 60° rotation of the amine group to form an N−H···S hydrogen bond, but the carbonyl stretch for **Cys-VIII** appears at 1794 cm^{-1}. It is therefore very surprising both that **Cys-II** and **Cys-VIII** have such different stretching frequencies, and also that **Cys-II**, one of the lowest energy conformers, is absent from the matrix.

The S−H stretch varies only over a narrow range of 2539−2560 cm^{-1}. This is consistent with the geometries since S−H seems to form only very weak hydrogen bonds. In general, the absorptions are also predicted to be quite weak, so that the S−H peak is not likely to be useful in distinguishing conformers.

3.5. Rotational Spectra. Alonso et al.[25] recently reported the identification of six low-energy conformers of Cys through Fourier transform microwave spectroscopy, providing a good opportunity here to compare the computed and experimental results. Computed rotational constants, centrifugal distortion constants, and dipole moments are reported in Table 4 along with experimental values, where available.

Table 4. Equilibrium {A_e, B_e, C_e} and Ground-State {A_0, B_0, C_0} Rotational Constants, Quartic Centrifugal Distortion Constants in the *A*-Reduced Representation, and Dipole Moments (μ_a, μ_b, μ_c) of the 11 Lowest-Energy Conformers of L-Cysteine[a]

	Cys-I	Cys-II	Cys-III	Cys-IV	Cys-V	Cys-VI	Cys-VII	Cys-VIII	Cys-IX	Cys-X	Cys-XI
A_e	3059.2	4376.2	2886.6	2855.8	4235.8	3094.6	3182.5	4534.7	4505.9	2977.2	2984.0
B_e	1641.1	1189.3	1654.3	1715.8	1195.0	1605.7	1607.4	1181.2	1185.4	1558.3	1610.3
C_e	1357.6	1028.6	1390.6	1430.2	1018.5	1335.4	1302.7	972.8	970.7	1234.9	1370.0
A_0	3039.4(3071.1)	4341.7(4359.2)	2852.1(2889.4)	2812.4	4204.9(4235.6)	3055.1	3174.6(3216.2)	4502.1	4462.2	2977.1(3004.2)	2953.9
B_0	1623.0(1606.5)	1180.1(1178.3)	1642.6(1623.0)	1698.8	1187.1(1187.3)	1601.9	1587.9(1572.7)	1172.0	1176.4	1538.7(1527.4)	1605.2
C_0	1344.5(1331.8)	1018.5(1015.3)	1383.6(1367.8)	1421.0	1007.5(1003.1)	1331.3	1287.8(1276.8)	965.2	963.2	1218.3(1210.7)	1360.5
D_{JK}	−434	898	−449	−891	1036	−1054	−2621	706	763	−1145	−1416
D_J	471	89	628	628	99	614	562	59	59	438	749
D_K	1757	835	2049	2014	1379	2809	9180	1028	1187	5316	5164
μ_a	−1.55	2.84	1.00	1.08	1.69	2.52	2.06	−2.56	−2.46	2.30	1.22
μ_b	4.08	−2.62	−1.34	0.73	−0.34	−4.94	0.47	2.83	3.80	−0.36	−1.84
μ_c	−1.27	1.16	1.45	2.51	−0.66	1.74	0.04	0.29	0.66	−0.24	0.18

[a] Rotational constants given in MHz and belong to the structures optimized at the MP2(FC)/aug-cc-pV(T+d)Z level. Quartic centrifugal distortion constants are given in Hz. Dipole moments are given in Debye and computed at the B3LYP/aug-cc-pVTZ level. Where available, experimental values[25] are given in parentheses. Ground-state rotational constants are determined from MP2(FC)/aug-cc-pVTZ equilibrium structures with anharmonic B3LYP/6-31G* corrections.

Some of the conformers (*e.g.*, **Cys-I**, **Cys-II**, and **Cys-VI**) have substantial dipole moments along the principal axes, thus helping (a) the observation of the related rotational transitions, and (b) the assignment of conformers based on information about which rotational constants correspond to intense transitions. Generally, we can divide the cysteine conformers into two groups based on the orientation about the $C_\alpha-C_\beta$ bond. In gauche conformers, the thiol is gauche to both the amine and carboxyl groups. The overall geometry is therefore more compact about the *B* axis, which is reflected in the larger B_e rotational constants for **Cys-I** and **Cys-III** in comparison to **Cys-II** and **Cys-V** (Table 4). In contrast, in a trans conformation, the thiol is positioned antiperiplanar to either the amine or carboxyl group. The overall geometry for the trans conformers is therefore extended along the *A* axis, leading to much larger A_e rotational constants for **Cys-II** and **Cys-V** in comparison to **Cys-I** and **Cys-III**.

At the MP2(FC)/aug-cc-pV(T+d)Z level, the rotational constants corresponding to the optimized equilibrium structures should be accurate enough to be useful to deduce the presence of conformers when interpreting experimental microwave spectra. As seen in Table 4, this is indeed the case. The mean absolute deviations from experiment are 15, 25, and 21 MHz for A_e, B_e, and C_e, respectively. Vibrational corrections for the rotational constants can be evaluated within the realm of second-order vibrational perturbation theory (VPT2)[60−63] by taking one-half the sum of the lowest-order vibration−rotation interaction constants α_i. The mean absolute deviations for B_0 and C_0 become only 11 and 9 MHz.

For the *A* axis, the theoretical rotational constants consistently underestimate the experimental ones by as much as 40 MHz. For most conformers, the cysteine molecule lies along the *A* axis, with the C−S bond running roughly perpendicular along the *C* axis. The moment of inertia about the *A* axis is greatly affected by the position of the sulfur atom, and A_0 is therefore very sensitive to the C−S bond length. If the C−S bond length is systematically overestimated, then the computed A_0 constants will be too small. For example, shortening the C−S bond length by 0.005 Å in **Cys-I** increases A_e by 30 MHz, bringing the corresponding A_0 into nearly exact agreement with experiment. Such bond length discrepancies may be attributed to a number of factors, including neglect of core correlation, basis set incompleteness, or higher excitations that would be included in CCSD or CCSD(T) geometry optimizations. The errors in B_0 and C_0 also appear to be systematic, with both being consistently overestimated. In all cases, the zero-point vibrational corrections lower the rotational constant, consistent with the vibrationally averaged bond lengths being longer than their equilibrium values. Zero-point vibrational corrections to the rotational constants therefore improve agreement for B_0 and C_0, but actually diminish the agreement for A_0. The source of the systematic error for B_0 and C_0 is more difficult to assess than for A_0, especially without the empirical refinement that was performed for conformers of glycine and proline.[14,17]

Table 5. Quadrupole Coupling Constants for Conformers of Cysteine Computed at the MP2/cc-pVTZ-LD Level (See Text)[a]

Conformer	χ_{aa}	χ_{bb}	χ_{cc}
Cys-I	−3.14 (−3.12)	2.44 (2.44)	0.70 (0.68)
Cys-II	−0.18 (−0.41)	2.19 (2.23)	−2.01 (−1.83)
Cys-III	−0.01 (−0.15)	0.34(0.44)	−0.32 (−0.30)
Cys-IV	−3.09	2.74	0.35
Cys-V	−4.39	2.74	1.65
Cys-VI	−3.26	2.36	0.9
Cys-VII	0.06 (0.00)	−0.48 (−0.45)	0.42 (0.45)
Cys-VIII	−3.02	1.51	1.51
Cys-IX	−3.22	1.61	1.61
Cys-X	0.51	−1.99	1.49
Cys-XI	−1.27	0.88	0.39

[a] Experimental values[25] where known are given in parentheses. All values given in MHz.

In the experiments of Alonso et al.,[25] some ambiguity still remained in differentiating conformers with similar rotational constants. For example, the B_0 and C_0 rotational constants of **Cys-I** and **Cys-III** match within 40 MHz while A_0 matches within 180 MHz. Furthermore, the computed rotational constants for **Cys-I** lie in between the observed values. For example, C_0 for **Cys-I** is computed to be 1344.5 MHz, in between the observed values of 1331.8 and 1367.8 MHz. Quadrupole coupling constants of the nitrogen nucleus are therefore necessary to uniquely identify such conformers. The quadrupole coupling constants ($\chi_{\alpha\alpha}$) are given as[88]

$$\chi_{\alpha\alpha} = eq_{\alpha\alpha}Q \tag{5}$$

where $q_{\alpha\alpha}$ is the electric field gradient along the α-axis at the nitrogen nucleus, *e* is the fundamental charge, and *Q* is the nuclear quadrupole moment. For the nitrogen quadrupole moment, we use the literature value of 20.44 mb.[89] As seen for ammonia (Table S6, Supporting Information), the accurate computation of electric field gradients at the nuclei presents a difficult theoretical problem. Similar difficulties hold for spin-dependent properties that depend on contact terms since the amplitude and shape of the wave function near the nuclei must be very accurately described.[90,91] In particular, Gaussian basis functions have the incorrect shape at the nuclei, so that extremely flexible and carefully designed basis sets are required for accurate results.

In general, double-ζ basis sets and Hartree−Fock methods are not flexible enough to yield good results for quadrupole coupling constants. Since the coupling constant depends only on the nitrogen nucleus, it is possible to use a locally dense basis on the nitrogen atom.[92] The combination of a cc-pCV5Z basis on nitrogen and cc-pVTZ basis on all other atoms (denoted cc-pVTZ-LD) very closely matches both the full cc-pCV5Z result and the experimental coupling constant (4.09 MHz)[93] for ammonia (Table S2, Supporting Information). Furthermore, probably through fortuitous cancelation of errors, MP2 matches the CCSD(T) results well, better than even CCSD. The combination of MP2 and the locally dense basis therefore seems to offer the best combination of accuracy and efficiency for computing the quadrupole coupling constants of cysteine.

The computed quadrupole coupling constants for the conformers of cysteine are presented in Table 5. The MP2/

cc-pVTZ-LD approximation generally performs quite well, yielding most coupling constants within 0.1 MHz of experiment with the largest deviation being 0.23 MHz. In particular, the ambiguity in assignment based on rotational constants is now removed. For example, **Cys-I** has a strong χ_{aa} quadrupole coupling while **Cys-III** exhibits almost no χ_{aa} coupling, in agreement with the experimental results of Alonso et al.[25]

4. Summary

In the present work, we performed a comprehensive study of the important structural features and spectroscopic signatures of the amino acid L-cysteine. Through the focal point approach, we have established definitive relative energies of the eleven lowest conformers to within a standard error of 0.3 kJ mol^{-1} (1σ) or 95% confidence interval of $\pm$ 0.6 kJ mol^{-1} (2σ).

Because of the added flexibility of the thiol side chain, cysteine exhibits 71 unique conformers (fully specified in Supporting Information) and eleven conformers within 10 kJ mol^{-1} of the lowest minimum. As observed previously,[1] Hartree−Fock energies are inaccurate. Inclusion of electron correlation with B3LYP greatly improves results, but still fails by more than 2.5 kJ mol^{-1} for some conformers, which becomes significant when so many conformers lie within a narrow 10 kJ mol^{-1} range. In general, B3LYP performs well for geometries and zero-point vibrational corrections, but inclusion of correlation through at least MP2 seems necessary for accurate energies. Definitive energies to 0.5 kJ mol^{-1} accuracy still require corrections through CCSD(T).

While hydrogen bonding and electrostatics are the most important factors determining structures and energetics, the bond length, bond angle, and energy changes between conformers depend strongly on subtle electronic effects, including hyperconjugation, steric repulsion, hydrogen-bond cooperativity, and dispersion forces. In contrast to previous work, we therefore emphasize features such as the trans angle rule[80] and the gauche effect.[84,85] An additive picture of hydrogen bonds may therefore be overly simplistic for cysteine.

Harmonic frequencies were computed at the B3LYP/aug-cc-pVTZ level with anharmonic corrections at the B3LYP/6-31G* level (Tables 3 and S4, Supporting Information). The vibrational perturbation theory generally performs well (within 20 cm^{-1} of experiment), although it breaks down for a few large-amplitude motions with very low frequencies. The computed fundamentals should aid future IR spectroscopy studies. Since we are aiming for accuracy near 0.5 kJ mol^{-1} in the conformational energies, rigorous anharmonic zero-point vibrational corrections are necessary instead of simply scaling harmonic frequencies.

The extensive ab initio results reported here should serve as an important reference both for calibrating more approximate theoretical methods or future experiments, including circular dichroism or infrared and microwave spectroscopy of isotopologues of cysteine. As more empirical data becomes available (e.g., rotational constants of isotopologues), the structures and energies can be further refined by empirical fitting, as done previously for glycine[14] and proline.[17]

Acknowledgment. This work was supported by NSF grant CHE-0749868, an NSF-MTA-OTKA grant, and the Hungarian Scientific Research Fund (OTKA, K72885, IN77954). Most computations were run at the Pittsburgh Supercomputing Center under TeraGrid grant TG-CHE070039N.

Supporting Information Available: Comparison of cysteine conformational energies with previously computed values, zero-point vibrational energy comparison for B3LYP/aug-cc-pVTZ and MP2/DZP++ levels of theory, summary of N−C−C bond angle and hyperconjugation changes due to amine rotation, NBO summary of hyperconjugation for varying conformations about the $C_\alpha-C_\beta$ bond, complete table of anharmonic vibrational fundamentals and double-harmonic infrared relative intensities, electron correlation and basis set dependence of computed quadrupole coupling constants of ammonia, optimized geometries and energies at MP2(FC)/aug-cc-pV(T+d)Z level of theory for eleven lowest energy conformers of L-cysteine, and optimized geometries and energies at MP2(FC)/cc-pVTZ level of theory for 71 unique conformers of L-cysteine. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Császár, A. G.; Perczel, A. *Prog. Biophys. Mol. Biol.* **1999**, *71*, 243–309.

(2) Cox, M. M.; Nelson, D. L., *Principles of Biochemistry*, 4th ed; W.H. Freeman and Company: New York, 2005.

(3) Lu, X. F.; Galkin, A.; Herzberg, O.; Dunaway-Mariano, D. *J. Am. Chem. Soc.* **2004**, *126*, 5374–5375.

(4) Li, H. M.; Thomas, G. J. *J. Am. Chem. Soc.* **1991**, *113*, 456–462.

(5) Császár, A. G. *J. Am. Chem. Soc.* **1992**, *114*, 9568–9575.

(6) Hu, C. H.; Shen, M. Z.; Schaefer, H. F. *J. Am. Chem. Soc.* **1993**, *115*, 2923–2929.

(7) Allen, W. D.; East, A. L. L.; Császár, A. G., *Structures and Conformations of Non-Rigid Molecules*; Kluwer: Dordrect, The Netherlands, 1993; p 343.

(8) Császár, A. G.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1998**, *108*, 9751–9764.

(9) East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638–4650.

(10) Császár, A. G.; Tarczay, G.; Leininger, M. L.; Polyansky, O. L.; Tennyson, J.; Allen, W. D. *Spectroscopy from Space*; Demaison, J., Sarka, K., Eds.; Kluwer: Dordrecht, The Netherlands, 2001; p 317−339.

(11) Gonzales, J. M.; Pak, C.; Cox, R. S.; Allen, W. D.; Schaefer, H. F.; Császár, A. G.; Tarczay, G. *Chem.—Eur. J.* **2003**, *9*, 2173–2192.

(12) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *120*, 11586–11599.

(13) Császár, A. G. *J. Mol. Struct.* **1995**, *346*, 141–152.

(14) Kasalová, V.; Allen, W. D.; Schaefer, H. F.; Czinki, E.; Császár, A. G. *J. Comput. Chem.* **2007**, *28*, 1373–1383.

(15) Császár, A. G. *J. Phys. Chem.* **1996**, *100*, 3541–3551.

(16) Szidarovszky, T.; Czakó, G.; Császár, A. G. *Mol. Phys.* **2009**, DOI: 10.1080/00268970802616350.

(17) Allen, W. D.; Czinki, E.; Császár, A. G. *Chem.—Eur. J.* **2004**, *10*, 4512–4517.

(18) Czinki, E.; Császár, A. G. *Chem.—Eur. J.* **2003**, *9*, 1008–1019.

(19) Laurence, P. R.; Thomson, C. *Theor. Chim. Acta* **1981**, *58*, 121–124.

(20) Wright, L. R.; Borkman, R. F. *J. Am. Chem. Soc.* **1980**, *102*, 6207–6210.

(21) Schäfer, L.; Kulpnewton, S. Q.; Siam, K.; Klimkowski, V. J.; Vanalsenoy, C. *J. Mol. Struct. (THEOCHEM)* **1990**, *68*, 373–385.

(22) Gronert, S.; O'Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, *117*, 2071–2081.

(23) Krishnan, R.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 4244–4245.

(24) Dobrowolski, J. C.; Rode, J. E.; Sadlej, J. *J. Mol. Struct. THEOCHEM* **2007**, *810*, 129–134.

(25) Sanz, M. E.; Blanco, S.; Lopez, J. C.; Alonso, J. L. *Angew. Chem., Int. Ed.* **2008**, *47*, 6216–6220.

(26) Frisch, M. J.; Pople, J. A.; Delbene, J. E. *J. Phys. Chem.* **1985**, *89*, 3664–3669.

(27) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.

(28) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.

(29) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

(30) Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

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

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

(33) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572–4585.

(34) Wilson, A. K.; Dunning, T. H., Jr. *J. Chem. Phys.* **2003**, *119*, 11712–11714.

(35) Pulay, P. *Mol. Phys.* **1969**, *17*, 197–204.

(36) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley-Interscience: New York, 1986.

(37) Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266.

(38) Crawford, T. D.; Schaefer, H. F. *Rev. Comp. Chem.* **2000**, *14*, 33–136.

(39) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

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

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

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

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

(44) Jayatilaka, D.; Lee, T. J. *J. Chem. Phys.* **1993**, *98*, 9734–9747.

(45) Taylor, P. R.; Lee, T. J. *Int. J. Quantum Chem. Symp.* **1989**, *23*, 199.

(46) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; Auer, A. A.; Bernholdt, D. B.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; O'Neill, D. P.; Price, D. R.; Ruud, R.; Schiffmann, F.; Varner, M. E.; Vázquez, J. *ACES II*, Mainz-Austin: Budapest, Hungary, 2005.

(47) Janssen, C. L.; Nielsen, I. B.; Leininger, M. L.; Valeev, E. F.; Seidl, E. T., *The Massively Parallel Quantum Chemistry Program (MPQC)*, version 2.3.1; Sandia National Laboratories: 2004.

(48) Nielsen, I. M. B. *Chem. Phys. Lett.* **1996**, *255*, 210–216.

(49) Nielsen, I. M. B.; Seidl, E. T. *J. Comput. Chem.* **1995**, *16*, 1301–1313.

(50) 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.; An P. Palmieri, A. N.; Pitzer, R.; Schumann, U.; Stoll, H.; A. J. Stone, R. T.; Thorsteinsson, T., *MOLPRO*, version 2006.1; 2006.

(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kuding, 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; 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.; Octerski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenber, J. J.; Zkrzewski, V. G.; Dappich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Forseman, 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.; Nanykkara, A.; Callacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A., *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.

(52) Eliel, E. L.; Wilen, S. H.; Mander, L. N., *Stereochemistry of Organic Compounds*; John Wiley & Sons: New York, 1994.

(53) Klopper, W.; Kutzelnigg, W. *J. Mol. Struct. (THEOCHEM)* **1986**, *28*, 339–356.

(54) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330–333.

(55) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104–6114.

(56) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(57) Tarczay, G.; Császár, A. G.; Klopper, W.; Quiney, H. M. *Mol. Phys.* **2001**, *99*, 1769–1794.

(58) Handy, N. C.; Yamaguchi, Y.; Schaefer, H. F. *J. Chem. Phys.* **1986**, *84*, 4481–4484.

(59) Allen, W. D.; Császár, A. G. *J. Chem. Phys.* **1993**, *98*, 2983–3015.

(60) Clabo, D. A., Jr.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F. *Chem. Phys.* **1988**, *123*, 187–239.

(61) Allen, W. D.; Yamaguchi, Y.; Császár, A. G.; Clabo, D. A., Jr.; Remington, R. B.; Schaefer, H. F. *Chem. Phys.* **1990**, *145*, 427–466.

(62) Nielsen, H. H. *Rev. Mod. Phys.* **1951**, *23*, 90–136.

(63) Mills, I. M., *Molecular Spectroscopy: Modern Research*; Rao, K. N., Mathews, C. W., Eds.; Academic Press: New York, 1972; p 115.

(64) Schuurman, M. S.; Allen, W. D.; Schleyer, P. v. R.; Schaefer, H. F. *J. Chem. Phys.* **2005**, *122*, 104302. Schuurman, M. S.; Allen, W. D.; Schaefer, H. F. *J. Comput. Chem.* **2005**, *26*, 1106–1112.

(65) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059–3065.

(66) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(67) Reed, A. E.; Weinhold, F.; Curtiss, L. A.; Pochatko, D. J. *J. Chem. Phys.* **1986**, *84*, 5687–5705.

(68) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1983**, *78*, 4066–4073.

(69) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211–7218.

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

(71) Simmonett, A. C.; Evangelista, F. A.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2007**, *127*, 014306.

(72) Zhang, X.; Maccarone, A. T.; Nimlos, M. R.; Kato, S.; Bierbaum, V. M.; Ellison, G. B.; Ruscic, B.; Simmonett, A. C.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2007**, *126*, 044312.

(73) Wilke, J. J.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2008**, *128*, 074308.

(74) Czakó, G.; Mátyus, E.; Simmonett, A. C.; Császár, A. G.; Schaefer, H. F.; Allen, W. D. *J. Chem. Theory Comput.* **2008**, *4*, 1220–1229.

(75) Simmonett, A. C.; Schaefer, H. F.; Allen, W. D. *J. Chem. Phys.* **2009**, *130*, 044301.

(76) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.

(77) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129–4141.

(78) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vazquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

(79) Dunning, T. H., Jr. *J. Chem. Phys.* **1970**, *53*, 2823–2833.

(80) Rasänen, M.; Aspiala, A.; Homanen, L.; Murto, J. *J. Mol. Struct.* **1982**, *96*, 81–100.

(81) Pross, A.; Radom, L.; Riggs, N. V. *J. Am. Chem. Soc.* **1980**, *102*, 2253–2259.

(82) Weinhold, F.; Landis, C. L., *Valency and Bonding*; Cambridge Unversity Press: Cambridge, UK, 2005.

(83) Allinger, N. L.; Fermann, J. T.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1997**, *106*, 5143–5150.

(84) Goodman, L.; Gu, H. B.; Pophristic, V. *J. Phys. Chem. A* **2005**, *109*, 1223–1229.

(85) Lesarri, A.; Cocinero, E. J.; Lopez, J. C.; Alonso, J. L. *J. Am. Chem. Soc.* **2005**, *127*, 2572–2579.

(86) Carbonniere, P.; Lucca, T.; Pouchan, C.; Rega, N.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384–388.

(87) Dobrowolski, J. C.; Jamroz, M. H.; Kolos, R.; Rode, J. E.; Sadlej, J. *ChemPhysChem* **2007**, *8*, 1085–1094.

(88) Hollas, J. M., *High Resolution Spectroscopy*, 2nd ed; Wiley: New York, 1998.

(89) Polak, R.; Fiser, *J. Chem. Phys.* **2008**, *351*, 83–90.

(90) Helgaker, T.; Jaszunski, M.; Ruud, K.; Gorska, A. *Theor. Chem. Acc.* **1998**, *99*, 175–182.

(91) Benedikt, U.; Auer, A. A.; Jensen, F. *J. Chem. Phys.* **2008**, *129*, 064111.

(92) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 6201–6208.

(93) Grigolin, P.; Moccia, R. *J. Chem. Phys.* **1972**, *57*, 1369–1376.

CT900005C

JCTC Journal of Chemical Theory and Computation

Balance of Attraction and Repulsion in Nucleic-Acid Base Stacking: CCSD(T)/Complete-Basis-Set-Limit Calculations on Uracil Dimer and a Comparison with the Force-Field Description

Claudio A. Morgado,[†,‡] Petr Jurečka,[†,§] Daniel Svozil,[†] Pavel Hobza,[‡,§] and Jiří Šponer*[,†,‡]

Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo náměstí 2, 166 10 Prague 6, Czech Republic, and Department of Physical Chemistry, Palacky University, tr. Svobody 26, 771 46, Olomouc, Czech Republic

Received January 7, 2009

Abstract: We have carried out reference quantum-chemical calculations for about 100 geometries of the uracil dimer in stacked conformations. The calculations have been specifically aimed at geometries with unoptimized distances between the monomers including geometries with mutually tilted monomers. Such geometries are characterized by a delicate balance between local steric clashes and local unstacking and had until now not been investigated using reference quantum-mechanics (QM) methods. Nonparallel stacking geometries often occur in nucleic acids and are of decisive importance, for example, for local conformational variations in B-DNA. Errors in the short-range repulsion region would have a major impact on potential energy scans which were often used in the past to investigate local geometry variations in DNA. An incorrect description of such geometries may also partially affect molecular dynamics (MD) simulations in applications when quantitative accuracy is required. The reference QM calculations have been carried out using the MP2 method extrapolated to the complete basis-set limit and corrected for higher-order electron-correlation contributions using CCSD(T) calculations with a medium-sized basis set. These reference calculations have been used as benchmark data to test the performance of the DFT-D, SCS(MI)-MP2, and DFT-SAPT QM methods and of the AMBER molecular-mechanics (MM) force field. The QM methods show close to quantitative agreement with the reference data, albeit the DFT-D method tends to modestly exaggerate the repulsion of steric clashes. The force field in general also provides a good description of base stacking for the systems studied here. However, for geometries with close interatomic contacts and clashes, the repulsion effects are rather severely exaggerated. The discrepancy reported here should not affect the overall stability of MD simulations and qualitative applications of the force field. However, it may affect the description of subtle quantitative effects such as the local conformational variations in B-DNA. Preliminary calculations for two H-bonded uracil base pairs, including one with a C−H···O H-bond, indicate excellent performance of the tested QM methods for all intermonomer distances. The force field, on the other hand, is less satisfactory, especially in the repulsive regions.

Introduction

Base stacking interactions provide a substantial part of the thermodynamics stability of nucleic acids, shape their structure, and contribute to their dynamics.[1−7] QM calculations with the inclusion of electron-correlation effects belong to important tools that help to understand the role of stacking interactions in nucleic acids.[8−18] Although the gas-phase QM calculations do not directly correlate with the thermodynamics stability of nucleic acids due to the complex interplay between molecular forces in solvated nucleic acids, they reveal direct and accurate structure-energy relationships which allow for an exhaustive description of the potential

* Corresponding author e-mail: sponer@ncbr.chemi.muni.cz.

[†] Institute of Biophysics, Academy of Sciences of the Czech Republic.

[‡] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic.

[§] Palacky University.

10.1021/ct9000125 CCC: $40.75
Published on Web 04/29/2009

energy surfaces (PES) of stacked structures of dimers of nucleic-acid bases.[19] Such calculations characterize the intrinsic forces between the stacked bases, are indispensable to reveal the nature of base stacking, and provide reference data to parametrize/verify other computational methods. Such calculations can also help to partially rationalize experimental data on nucleic acids, albeit direct transferability of the QM data to nucleic acids would require, e.g., inclusion of solvent screening of the gas-phase electrostatics which is specific for distinct nucleic-acid architectures.[20,21] The QM description of base stacking in general requires application of electron-correlation techniques and diffuse-polarization basis sets of atomic orbitals. The present golden standard for base-stacking computations is the CBS(T) method,[19] which is based on MP2[22] calculations with complete basis-set (CBS) extrapolation corrected for higher-order electron-correlation contributions using the CCSD(T)[23] Cizek method[24] with a medium-size basis set. In the present paper we have carried out an extensive QM characterization of stacked conformations of the uracil dimer, which is the smallest base stacking system found in nucleic acids. For the purpose of comparison we have also investigated two H-bonded dimers.

Uracil-Dimer Stacking in Nucleic-Acid Structures. The stacking of uracil bases can be seen in a wide variety of RNA structures, spanning a large range of different conformations. Besides the intrastrand and interstrand uracil dimers occurring in canonical 5′-UU-3′, 5′-AU-3′, and 5′-UA-3′ A-RNA base-pair steps, many diverse stacking arrangements between two uracil bases are observed in the experimental structures of RNA hairpins,[25,26] pseudoknots,[27,28] dimeric RNA quadruplexes,[29] SRE RNA,[30] U2 snRNA stem I from S. Cerevisiae,[31] a satellite tobacco mosaic virus/RNA complex,[32] and many other RNA systems including obviously the ribosomal structures.[33,34] Nevertheless, the main aim of this paper is not to investigate the stacking of uracil bases in particular RNAs but to better describe its physicochemical nature and assess the ability of other methods to evaluate stacking interactions, including a standard MM force field.

Preceding Theoretical Studies. The first electron-correlation characterization of the uracil dimer was reported by Šponer et al.[35] using the MP2/6-31G*(0.25)[36,37] method. The study concluded that the stability of stacking originates in the dispersion attraction and that the orientation dependence stems from the electrostatic term. The calculations ruled out several incorrect models of stacking and showed a surprisingly good performance of common molecular-mechanics force fields[38] combining Lennard-Jones potentials with Coulombic terms with atom-centered point charges. Kratochvíl et al.[39] investigated the potential- and free-energy surfaces of the uracil dimer in the gas phase, employing a combination of *ab initio*, empirical potential, computer simulations, and statistical thermodynamics techniques. They reported eleven low-energy minima in the PES of the uracil dimer: seven H-bonded, one T-shaped, and three stacked structures. The global gas-phase free-energy minimum was predicted to be an H-bonded structure, while stacked structures were found to be less populated than H-bonded ones. N1-methylation shifted the free energy balance in favor of the stacked structures, due to the increased dispersion energy and elimination of several stable H-bonded structures.[40] Hobza and Šponer[41] carried out gas-phase gradient optimizations of a stacked uracil dimer that revealed substantial deformations of the monomers in the gas-phase complexes. The free energy of stacking of the uracil dimer in water was estimated by Florian et al.[42] using MP2/6-31G*(0.25) and Langevine-dipole calculations, concluding that the electrostatic component of stacking, which determines the mutual orientation of bases in the gas phase, is eliminated by solvent screening effects. Later, Leininger et al.[43] and Hobza and Šponer[44] reported the first large-scale MP2 calculations on the uracil dimer supplemented with CCSD(T) corrections, which were in meaningful agreement with the preceding MP2/6-31G*(0.25) data. Very recently Czyżnikowska et al.[45] studied the rise and twist dependence of the interaction energy components for the undisplaced (the centers of mass stacked one above the other) uracil dimer at the MP2/aug-cc-pVDZ[46,47] level of theory, using a variational-perturbational decomposition scheme. These calculations confirmed that the second-order dispersion term is basically independent of the twist angle, while the first-order electrostatic term shows a strong angular dependence. Cybulski and Sadlej[48] characterized an H-bonded and a stacked uracil dimer using the SAPT[49] and SAPT(DFT)[50] methods and demonstrated that the ratio of the dispersion term to the total interaction energy clearly differentiates between H-bonding and stacking interactions. So far the most comprehensive methodological study on the interaction energy of the uracil dimer is the work of Pitoňák et al.,[51] who employed a number of QM methods to evaluate the intermolecular interaction on the H-bonded and stacked structures of the dimer reported in the S22 set[52] of reference geometries. The predicted stacking energy of -9.77 kcal mol^{-1}, obtained by combining MP2/[aug-cc-pVTZ→aug-cc-pVQZ] and CCSD(T)/[aug-cc-pVDZ→aug-cc-pVTZ] CBS extrapolations, is in very good agreement with the CBS(T) energy reported for the same stacked uracil dimer in the S22 set. This suggests convergence of the computations. Their study also showed that a good estimate of the ΔCCSD(T) term can already be obtained with relatively small basis sets, such as 6-31+G**, and that the DFT-D method of Jurečka et al.[53] agrees well with estimated CCSD(T)/aug-cc-pVTZ data along the entire potential energy curve (PEC) for both the H-bonded and stacked structures.

Available Experimental Data. The only experimental data on the stabilization energy of the uracil dimer in vacuo was obtained almost thirty years ago by Yanson et al.[54] The authors reported for the 1-methyluracil dimer a stabilization enthalpy of 9.5 kcal mol^{-1}, measured in a range of temperature of about 295−318 K using mass-field spectrometry. The structures that were present in these unique experiments, however, are not known. Standard QM computations cannot be directly compared with such experiments since they calculate 0 K interaction energies as the difference between pure electronic energies. The experimental stabilization enthalpy would have to be compared to the weighted average of stabilization enthalpies of all populated structures at the experimental conditions.[55] Another experimental work relevant to the uracil dimer was done by Casaes et al.,[56] who

measured the gas-phase infrared spectra of jet-cooled uracil clusters, thymine clusters, and uracil·water clusters. The authors found evidence for the presence of several double hydrogen-bonded uracil dimers and for the formation of a larger highly symmetrical cluster. Interestingly, no evidence was seen for T-shaped and stacked structures, which does not support the predictions made by Kratochvíl et al.[39] Regarding experimental data in the condensed phase, an estimate of vertical stacking interactions of uridine in aqueous solution can be found in the work of Ts'o et al.,[57] who reported for this nucleoside a free energy of association of 290 cal mol^{-1}. This number cannot be directly compared to gas-phase calculations either, because it includes the entropy cost of bringing the monomers together and many other terms related to solvation.

The Scope of the Present Study. The present study is aimed differently than the preceding ones. We have shown that AMBER, the leading molecular-mechanics force field for nucleic acids, provides a surprisingly good description of base stacking.[35] However, this does not mean that the force field provides an exact description of stacking. In fact, the MM force fields, albeit often providing a very insightful description of nucleic-acid structure and dynamics, are known to have limitations. Besides the overall topologies, nucleic-acid structures and functions are affected by subtle structural details. In canonical double helices these variations are known as local conformational variations, i.e., modest deviations from the average helices that are determined by base sequence and other molecular interactions, such as those caused by protein or drug binding, or crystal packing forces in X-ray experiments.[58−61] Accurate local positioning of bases is also important elsewhere, for example in the catalytic centers of ribozymes.[62,63] Local conformational variations are assumed to be primarily caused by base stacking forces and are of primary importance for indirect readout of proteins, sequence-dependent DNA elasticity, etc. Studies, experiments as well as theory, of local conformational variations turned out to be very difficult, as local conformational variations are associated with very subtle energy changes. Computational studies ranging from extensive analysis of the base-stacking PESs up to full-scale MD simulations could provide insights into the sequence-dependence of NA structure.[64−72] However, such quantitative studies would require an exceptionally high accuracy of the energy description of direct base−base interactions, solvent screening effects, and the conformational space of the sugar−phosphate backbone. Regarding base stacking, local conformational variations are often associated with interactions involving nonparallel (mutually tilted) bases, where close contacts (steric clashes) between nucleobase edges or exocyclic functional groups are combined with local unstacking in other parts of the stacked base-pair steps. Steric effects associated with the amino groups of guanine in the minor groove of CpG B-DNA steps[59,69] or helical-twist/base-pair-roll redistribution in alternating pyrimidine-purine A-RNA sequences belong to the best documented examples.[68,73,74] Nonparallel stacked bases are obviously very common in complex noncanonical RNA regions. When the bases are not parallel, the stacking geometry typically reflects a competition between a segment of the dimer where the monomers are locally unstacked and another segment where the monomers are sterically clashing. An accurate description of this local compression (clash) is important for a proper description of the whole stacked system. We have recently investigated several geometries of the CpG B-DNA steps using the CBS(T) reference method, and these calculations indeed suggested that once geometries with nonparallel bases are considered, the differences between the benchmark calculations and the force field (as well as other methods) can be significant.[75] For a proper description of local conformational variations the accuracy of the van der Waals (vdW) term of the force field is critical, as it determines the balance between the steric clashes and the partially unstacked regions. The electrostatic part of the stacking energy with its r^{-1} dependence is not contributing significantly to the energy changes associated with the steric contacts. In addition, the electrostatic components of stacking in DNA are in general effectively attenuated due to solvent screening. The solvent screening actually limits the direct applicability of gas-phase QM calculations in studies of DNA local conformational variations, because the dominant role of electrostatics for the gas-phase stacking-energy dependence on helical twist vanishes in solvated nucleic acids.

Herein we report an interaction-energy analysis of the PES of the uracil dimer in stacked conformations, covering a wide range of about 100 structures with specific emphasis given to geometries of dimers in tilted conformations and with unoptimized vertical separation. The reference points are obtained with the CBS(T) method. We first scan a four-dimensional space considering twist between parallel bases, displacements in x and y directions, and the vertical separation between the monomers. Then, we investigate several geometries with tilted bases, i.e., structures with competing clashing and unstacking. While purely compressed or extended dimers with parallel bases cannot occur in real nucleic-acid structures, structures with nonparallel bases can be accompanied with close interatomic contacts (steric clashes) and local unstacking even upon the overall optimization of the vertical separation between the bases or base-pair steps.[76] The reference CBS(T) calculations serve as a benchmark for several other computational methods: the AMBER force field,[38] DFT-D,[53] SCS(MI)-MP2,[77] and DFT-SAPT.[78] The choice of the AMBER force field is dictated by the need to evaluate the performance of its nonbonded empirical potential—whose vdW term is very similar to the vdW term of most other MM force fields—in a wider region of the PES of nucleic-acid base dimers. The DFT-D and the SCS(MI)-MP2 methods have been chosen because they have proven a cost-efficient way for obtaining interaction energies in good agreement with CCSD(T) or CBS(T) benchmarks. However, both methods have been parametrized against the S22 training set, which only contains noncovalent complexes optimized to a minimum, and their performance in regions other than the minimum has not been as widely tested. We have performed DFT-SAPT energy calculations and decompositions for all of the structures in order to gain some insight into the nature of the interaction.

Among other results we show that the MM force-field calculations are basically capable of providing a satisfactory description of base stacking for the uracil dimer. However, the agreement between the force field and the reference CBS(T) calculations breaks down in the repulsive regions of the PES. The observed differences are caused by the vdW term of the force field and are large enough to substantially affect the description of the fine local conformational variations in nucleic-acid duplexes. Therefore, the correct description of such geometries may represent one of the challenges for future refinements of MM force fields. The DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods are mutually much more consistent in the repulsive region, although the differences between them also increase upon incrementing the repulsion. We have also investigated the dependence of the interaction energy on the intermonomer distance for two H-bonded uracil dimers. The mutual agreement between the QM methods is surprisingly good. On the contrary, the discrepancy between the AMBER force field and the QM methods for short intermonomer distances is much more accentuated, with the force field severely exaggerating the repulsion as the H-bonding distance is decreased. Nevertheless, it is important to point out that, considering the reliability of molecular modeling, the inaccuracy of the force-field description for the H-bonded base pairs is not as painful as the differences reported for stacking. The H-bonded base pairs, in contrast to stacking, usually represent interactions that are well separated from the other interactions. Thus the exaggeration of short-range repulsion in base pairs is likely to lead to mere overestimation of H-bond lengths with no significant effects on the local conformational variations. We would like to note here that our present study focuses on the stacking interactions and that the H-bonded data are more limited. The balance of forces in H-bonded base pairs is very different from that in the stacked structures. This topic requires a thorough analysis, and work in this direction is currently underway. For the information concerning the accuracy of the DFT-D methods for both H-bonded and stacked complexes we refer the reader also to studies that are available in the literature. For instance, Sherrill[79] showed that Grimme's DFT-D (PBE-D[101]) performs very well for both H-bonded and stacked complexes. A feeling about the spin-component-scaled methods can be obtained from refs 13 and 80. Finally, the DFT-SAPT method has been shown to give highly accurate results for both H-bonded and π-bonded complexes,[120] and the accuracy of the intermolecular components has also been studied for both cases.[48]

Computational Details

Geometries. The structures have been obtained in the following way. First, a uracil monomer is optimized using the RI-MP2[81−83] method along with the cc-pVTZ[46] basis set. Then, it is placed in the xy plane (z=0) with the center of mass coinciding with the origin. The N1−H1 "glycosidic" bond is parallel to the y-axis and is pointing to the direction of negative y values, and the Watson−Crick face of the base is oriented toward the direction of negative x values (Figure 1), with a minor modification for the structures derived to study interatomic clashes (NP1-NP4, see below and Supporting Information). This first monomer is always fixed, and the second monomer is initially superimposed on the first one. Then the position of the second monomer is determined via 6 independent parameters. There are three angles that are applied counterclockwise and consecutively via rotational matrices: rotation around the x axis (γ), rotation around the y axis (α), and rotation around the z axis (ω). The first two angles are applied to introduce the tilting between bases. When comparing to B-DNA conformational parameters, the α, γ, and ω angles are roughly analogous to propeller twist, buckle, and helical twist angles. Then, the second monomer is shifted along the x and y axes by the parameters Δx and Δy. Finally, the vertical distance is adjusted by Δz. This means that in all cases Δz is equal to the distance between the center of mass of the second monomer and the xy plane, where the first monomer is located (note that $\Delta z \equiv r$ in the figures of the stacked dimers). As the stacking energy is very sensitive to vertical compression or extension of the dimer and the vertical dependence of the stacking energy reflects the balance of vdW interaction terms,[76] we have scanned the stacking-energy dependence on Δz for most combinations of α, γ, ω, Δx, and Δy.

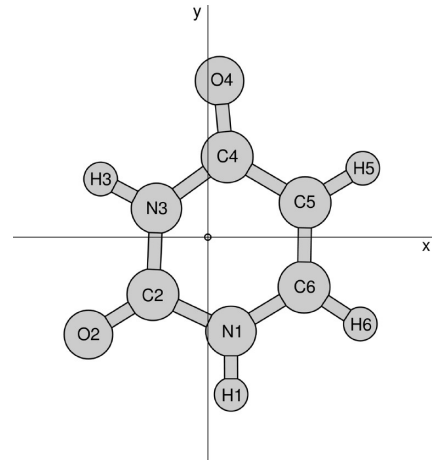


Figure 1. Orientation of the first uracil monomer in the xy plane.

The structures are grouped into two different sets, P and NP. The set P contains 9 different uracil dimers ($P\bar{1}$, P2, ..., P9), in conformations where the planes of the rings are parallel (P) to each other (α, $\gamma = 0$), whereas the set NP contains 4 different dimers (NP1, NP2, NP3, and NP4) in conformations in which one monomer is tilted with respect to the other one, that is, nonparallel (NP) conformations. A dimer is defined by those structural parameters that are kept fixed during the scan, where a free parameter is varied in order to build a potential-energy curve for each dimer. This way we have generated a subset of structures for each dimer, and these subsets add up to a total of 105 distinct geometries. With one exception ($P\bar{3}$, for which we vary the twist angle) we have scanned the Δz dependence of the stacking energy for a fixed combination of the other five geometrical parameters.

The four tilted dimers NP1-NP4 have been selected manually based on visual inspection of a number of α, γ, ω, Δx, and Δy combinations in order to obtain four vertical scans with diverse clashes. Obviously, one could imagine

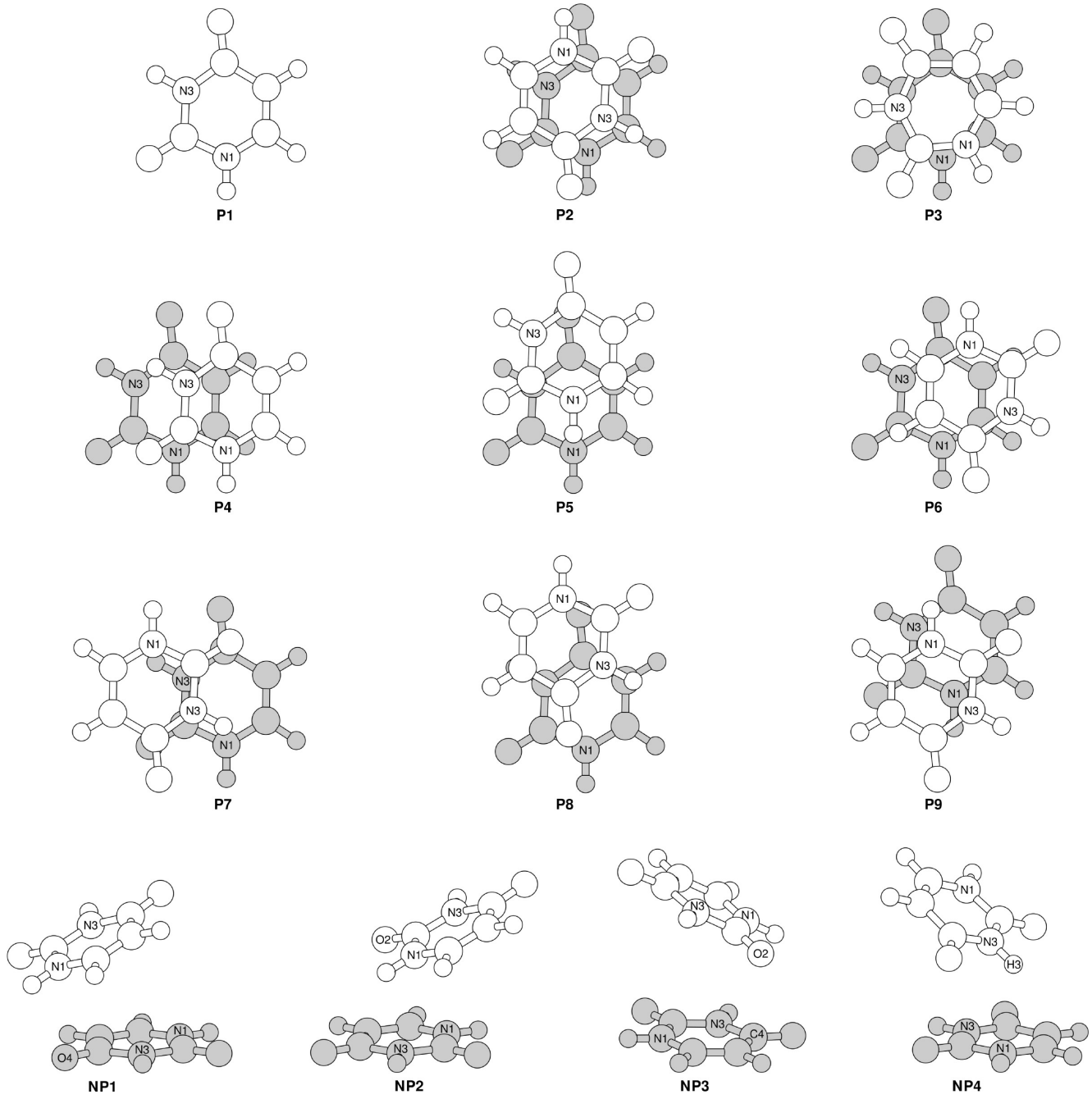


Figure 2. Top and side views of the parallel and nonparallel structures of the uracil dimer, respectively. The first uracil monomer is shown in gray. Some atoms have been labeled in order to aid in the visualization.

many other such geometries that would be worthy of study; however, the calculations remain expensive, and thus we have limited the number of studied structures.

For the H-bonding calculations the WC and "Calcutta" dimers have been selected (denoted HB1 and HB2, respectively), and the scans have been carried out by varying the O2···N3 and the O4···N3 distances, respectively. Note that the Calcutta base pair contains a C−H···O H-bond. The initial (equilibrium) geometries of both H-bonded dimers are taken from ref 84.

The Cartesian coordinates of all structures are given in the Supporting Information, while Figures 2 and 3 show the geometrical arrangement of the stacked and H-bonded dimers respectively.

Interaction Energies. The interaction energy ΔE^{AB} of a stacked or H-bonded complex is calculated according to the supermolecular approach as the difference in electronic energy between the complex and the isolated monomers (eq 1).[85,86] Since we have used only rigid monomers, we do not consider monomer deformation energies. In the case of the SCS(MI)-MP2 and CBS(T) methods, all of the interaction energies have been corrected for the basis set superposition error (BSSE) using the counterpoise (CP) technique.[87,88] This correction is not applied within the DFT-D formalism as this method was parametrized with respect to BSSE-corrected data.[53]

$$\Delta E^{AB} = E^{AB} - E^{A} - E^{B} \quad (1)$$

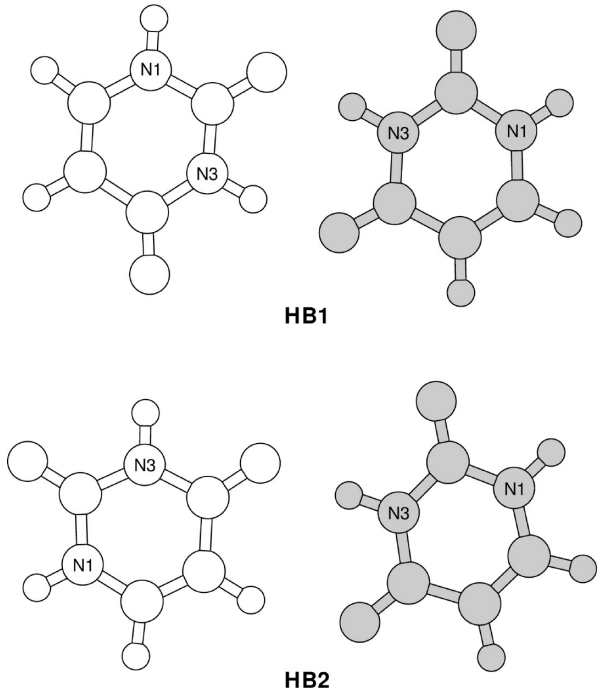


Figure 3. Top views of the H-bonded structures of the uracil dimer. Some atoms have been labeled in order to aid in the visualization.

The supermolecular approach is applied for all of the methods used in the present work, except for DFT-SAPT, where the interaction energy is given by the sum of selected perturbation contributions. In an intermolecular perturbation theory like DFT-SAPT the BSSE takes place only in the δHF term (see below).

Methods

AMBER Nonbonded Empirical Potential. The empirical-potential calculations of the interaction energies have been carried out with a local code, employing the vdW and Coulombic terms of the AMBER force field.[38] The atom-centered point charges were derived by means of an electrostatic potential (ESP) fitting at the MP2/aug-cc-pVDZ level of theory using the Merz−Kollman[89,90] methodology as implemented in the Gaussian 03[91] software package. This nonbonded potential is analogous to that in AMBER, with the only exception that condensed-phase simulations with AMBER are performed with charges derived in a RESP[92] fitting with the Hartree−Fock (HF) method and the 6-31G* basis set. Given that the HF method overestimates dipole moments, the use of HF charges implicitly accounts for polarization.

The choice of the MP2/aug-cc-pVDZ ESP charges instead of the original AMBER HF/6-31G* RESP charges deserves an explanation. One of the main purposes of this paper is to validate the physical correctness of the nonbonded potential of the AMBER force field. As demonstrated in the literature, meaningful benchmarking of the force field requires that the force-field electrostatic term is derived at the same computational level as used in the reference QM calculations.[35] Then the force-field electrostatic parameter set mimics as close as possible the respective Coulombic term of the full QM calculation. When using standard HF/6-31G* AMBER charges, we would inevitably bias the force field vs QM comparison. It would be often difficult to judge whether the observed differences reflect the basic limitation of the force-field model, the presence of interactions not included in the force-field model, or just the fact that the force-field and the QM calculations consider uncorrelated and correlated electrostatics, respectively. In addition, the present paper is focused on evaluating the capability of the van der Waals term of the force field to describe steric clashes. When introducing steric clashes, the electrostatic energy is basically constant. Note that a comparison of HF and MP2 ESP charges for base stacking is available in the literature.[93,94] In addition, in one of our earlier stacking papers we have also demonstrated that the use of out-of-plane charges or distributed multipoles does not improve the performance of the force field for base stacking compared with the set of atom-centered ESP-derived point charges.[95] Thus, the use of atom-centered MP2/aug-cc-pVDZ charges minimizes the difference between the force-field and QM description caused by the ESP electrostatics, allowing unambiguous assessment of the van der Waals part of the force field.

DFT Augmented with an Empirical Dispersion Term (DFT-D). The DFT-D calculations have been performed with the TurboMole 5.8[96] software package, in combination with a local code that computes the empirical dispersion correction. For the DFT part of the calculation we have used the RI approximation,[97−100] also known as density fitting. In the DFT-D approach as developed by Jurečka et al.,[53] a pairwise additive potential of the form $C_6 r^{-6}$ is used to account for long-range dispersion effects that are poorly described with common density functionals. The dispersion-corrected energy is given by

$$E_{DFT-D} = E_{KS-DFT} + E_{disp} \tag{2}$$

where $E_{KS\text{-}DFT}$ is the self-consistent Kohn−Sham energy, and E_{disp} is a term containing the empirical dispersion correction:

$$E_{disp} = -\sum_{ij} f_{damp}(r_{ij}, R^0_{ij}) C_{6,ij} r_{ij}^{-6} \tag{3}$$

$$f_{damp} = \frac{1}{1 + e^{-d(r_{ij}/(s_R R^0_{ij}) - 1)}} \tag{4}$$

In eq 3 r_{ij} is the distance between atoms i and j, R^0_{ij} is the equilibrium vdW separation between atoms i and j (derived from the atomic vdW radii), and $C_{6,ij}$ is the composite dispersion coefficient for the pair of atoms i and j (calculated from the corresponding atomic C_6 coefficients). The damping function (eq 4) is needed because the r^{-6} form is not valid at short distances, and because some short-range correlation effects are already present in the density functional. In this equation d is a parameter determining the steepness of the damping function, and s_R is a scaling coefficient that adjusts the magnitude of the vdW radius and that has been determined for several density-functional/basis-set combinations. The values of the C_6 atomic coefficients have been taken from the work of Grimme,[101] whereas the d and s_R parameters as well as the combination rules for the vdW

radii and the composite dispersion coefficients are the same as those in the work of Jurečka et al.[53] Here we use the TPSS[102] functional along with the 6-311++G(3df,3pd)[36] basis set. This level of theory complemented with the dispersion correction will hereafter be referred to as DFT-D.

SCS-MP2 for Molecular Interactions (SCS(MI)-MP2). The SCS(MI)-MP2[77] calculations have been performed with the Molpro 2006.1[103] software package, applying the frozen-core and density-fitting[104,105] approximations. This method is a reparameterization of the original SCS-MP2[106] method, in which the same- (SS) and opposite-spin (OS) components of the MP2 energy are empirically scaled ($E_{MP2} = c_{OS}E_{OS} + c_{SS}E_{SS}$) in an attempt to overcome the deficiencies of the MP2 theory, like the overestimation of the dispersion contribution to the correlation energy. The original SCS-MP2 method reduces the overestimation of the dispersion energy for stacked structures and thus provides very good estimates of the stabilization energy for these systems. Unfortunately, the method also reduces the dispersion for H-bonded structures, which results in this case in an underestimation of the stabilization energy. Using multivariate linear least-squares analysis and the CCSD(T) data in the S22 training set,[52] Distasio and Head-Gordon[77] found the optimal parameters that minimized the error between SCS-MP2 theory and CCSD(T), for the cc-pVXZ (X=T, Q) and extrapolated cc-pV(XY)Z (XY=DT, TQ) levels. The resulting method, known as SCS(MI)-MP2, provides very good estimates of stabilization energies for both planar H-bonded and stacked structures. Here we calculate the SCS(MI)-MP2 interaction energies using the cc-pV(DT)Z extrapolation (cc-pVDZ→cc-pVTZ), for which the same- and opposite-spin optimized parameters are 1.46 and 0.29, respectively. The MP2/CBS limit is approximated according to the following extrapolation scheme

$$E_{XY} = E_{SCF,Y} + \frac{X^3 E_{CORR,X} - Y^3 E_{CORR,Y}}{X^3 - Y^3} \quad Y > X \quad (5)$$

where $X = 2$ and $Y = 3$ for the D→T extrapolation used in this work.

DFT-Symmetry Adapted Perturbation Theory (DFT-SAPT). The DFT-SAPT[107−110] calculations have been performed with the Molpro 2006.1 software package. DFT-SAPT is a method that uses molecular properties from density functional theory in order to calculate intermolecular interaction energies by means of symmetry-adapted perturbation theory (SAPT). In this method the interaction energy is given as the sum of the first- and second-order energies, plus the δHF term. The first-order energy includes the electrostatic and exchange-repulsion contributions, while the second-order energy includes the induction, exchange-induction, dispersion, and exchange-dispersion contributions. The δHF term is an estimate of higher-order Hartree−Fock contributions and is determined as the difference of the HF interaction energy and the sum of the first- and second-order contributions, with the exception of the dispersion and exchange-dispersion energies. Since the HF interaction energy is calculated with BSSE-corrected monomer energies, the δHF term is BSSE dependent. The interaction energy is given by eq 6, and the electrostatic, induction, dispersion, and exchange contributions are defined in eqs 7−10.

$$E_{\text{int}} = E_{pol}^{(1)} + E_{ex}^{(1)} + E_{ind}^{(2)} + E_{ex-ind}^{(2)} + E_{disp}^{(2)} + E_{ex-disp}^{(2)} + \delta HF \quad (6)$$

$$E_{elec} = E_{pol}^{(1)} \quad (7)$$

$$E_{ind} = E_{ind}^{(2)} + E_{ex-ind}^{(2)} \quad (8)$$

$$E_{disp} = E_{disp}^{(2)} + E_{ex-disp}^{(2)} \quad (9)$$

$$E_{exch} = E_{ex}^{(1)} \quad (10)$$

For the energy decomposition we have employed the LPBE0AC[110] XC potential with the pure ALDA kernel for both the static and dynamic response, along with the aug-cc-pVDZ basis set. We have also utilized the density-fitting[110] approximation. The ionization potential (IP) of the monomer and the energy of the highest occupied molecular orbital (HOMO), which are required to evaluate the shift parameter, have been calculated at the PBE0[111]/aug-cc-pVDZ level of theory.

The relatively small aug-cc-pVDZ basis set has been used because our primary goal is to get an idea about the relative magnitude of the interaction energy components. DFT-SAPT interaction energies are usually underestimated when using this basis set, mainly because the dispersion component is underestimated by about 10−20%.[110] This underestimation does not change the conclusions regarding the relative importance of the individual contributions to the interaction energy given below. With a larger basis set DFT-SAPT was shown to provide very accurate total interaction energies in good agreement with the most accurate CCSD(T) calculations.[110] Using the aug-cc-pVDZ basis set for the IP calculations does not affect the shift values dramatically, and resulting errors should be smaller than the SAPT basis-set-size errors.

MP2/CBS Corrected for Higher-Order Correlation Effects (CBS(T)). The MP2 and CCSD(T) calculations have been performed with the Molpro 2006.1 software package, applying the frozen-core approximation. In the case of MP2, we have also applied the density-fitting[104,105] approximation. We have carried out the MP2/CBS calculations by extrapolating the Hartree−Fock and the correlation energies separately, following the Helgaker et al.[112,113] extrapolation scheme (eqs 11 and 12), in which E_X is the energy for the basis set with the largest angular momentum X, E_{CBS} is the energy for the complete basis set, and α is a parameter that was fitted in their original work. Herein we have used the aug-cc-pVDZ→aug-cc-pVTZ extrapolation.

$$E_X^{HF} = E_{CBS}^{HF} + Ae^{-\alpha X} \quad (11)$$

$$E_X^{CORR} = E_{CBS}^{CORR} + BX^{-3} \quad (12)$$

Given that higher-order correlation-energy contributions cannot be neglected, we have approximated the CCSD(T)/CBS interaction energies according to the following scheme:

$$\Delta E_{CBS}^{CCSD(T)} = \Delta E_{CBS}^{MP2} + (\Delta E^{CCSD(T)} - \Delta E^{MP2})\Big|_{6-31+G^{**}} \quad (13)$$

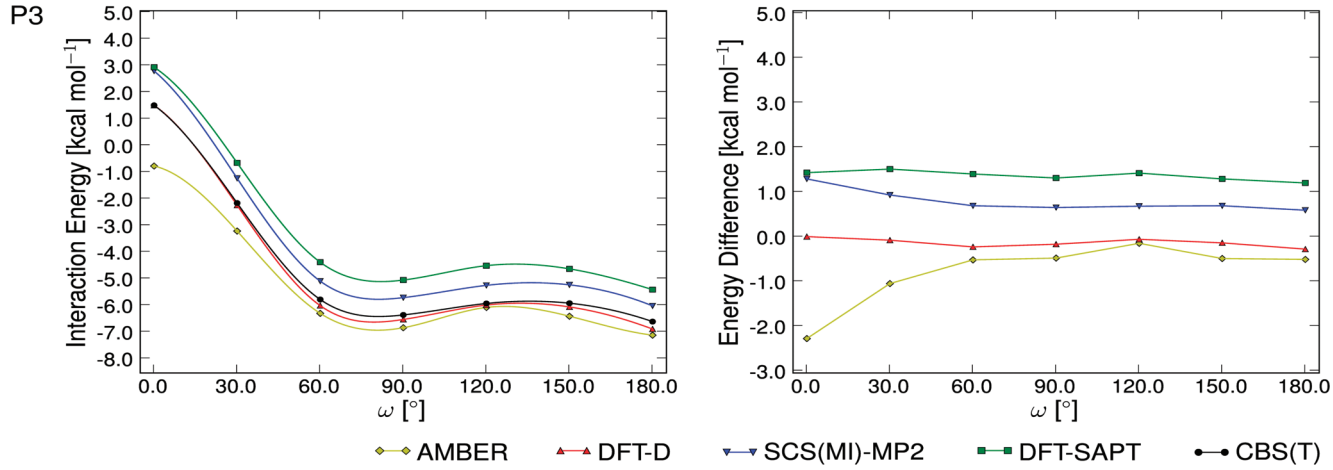


Figure 4. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for P3, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods. The vertical separation between the monomers is 3.3 [Å].

The use of eq 13 is based on the assumption that the difference between the CCSD(T) and MP2 interaction energies is basically basis-set independent and can therefore be evaluated with a small- or medium-size basis set. This assumption was validated for several H-bonded[114] and stacked[44] clusters, including H-bonded and stacked structures of the uracil dimer.[115] A CCSD(T)/CBS interaction energy estimated according to eq 13 is also known as CBS(T).[19]

Data Analysis. The interaction-energy data have been analyzed in terms of root-mean-squared (rms) errors, maximum absolute deviations (MAX), and XY plots. rms errors are calculated using the standard formula. In this work we define a deviation as the difference between the value obtained with a given method (AMBER, DFT-D, SCS(MI)-MP2, or DFT-SAPT) and the value obtained with CBS(T), and the MAX value for a given dimer simply corresponds to the absolute value of the largest deviation. All of the plots have been generated with the Matplotlib[116] graphics package in combination with the SciPy[117] and NumPy[118] packages. Energies and plots are given in the Supporting Information.

Results and Discussion

The Overall Performance of the Methods. Let us first assess the data qualitatively. Figure 4 shows the dependence of the stacking energy on twist at $r = 3.3$ Å (P3), which is primarily determined by the electrostatic term. The twist energy variation is rather modest due to the lower polarity of uracil compared to cytosine and guanine.[35] The DFT-D method achieves a close to exact match to the CBS(T) curve. The AMBER force field is more stabilizing than CBS(T), and also the profile of the twist dependence somewhat deviates from the CBS(T) data. The SCS(MI)-MP2 and DFT-SAPT methods systematically underestimate the binding, while the shapes of the curves basically match the CBS(T) one. In the case of the DFT-SAPT method the underestimation is mainly due to the small basis set used (aug-cc-pVDZ); with a larger basis set yet closer match with the CBS(T) data is expected.

Figure 5 shows the vertical scans for P1 and P2, i.e., the untwisted and antiparallel undisplaced dimers (untwisted and antiparallel refer to ω angles of 0 and 180°, respectively). When assessing the methods using the vertical scans, the most important region is that ~0.2−0.3 Å around the CBS(T) minimum, while the most relevant descriptor is the slope (gradient) of the PEC at a given distance within the repulsive region. This is related to the force associated with the repulsion. Two findings are apparent. First, all methods show very flat energy dependence around the vertical-separation minimum for the untwisted P1 dimer, which has the most repulsive electrostatic arrangement. The AMBER force field gives the shortest optimal vertical separation of the monomers for this arrangement. In contrast, the force field has an excessively steep onset of the repulsion in the short-range repulsion region of the P2 dimer which is also associated with the overestimation of the optimal intermonomer distance. The data for the P2 dimer also suggest that DFT-D is little overestimating the repulsion in the short-separation region, while SCS(MI)-MP2 provides a curve with almost the same slope as the CBS(T) one, albeit the method underestimates the interaction energy. DFT-SAPT also underestimates the interaction energy.

Figure 6 shows the vertical scans for the two untwisted-displaced dimers (P4 and P5). These are still in the electrostatically repulsive orientation, but the geometrical overlap of the bases is reduced. The most visible result is the overestimation of the short-range repulsion and of the optimal intermonomer distance by the force field.

Figure 7 summarizes the vertical scans on the four antiparallel displaced dimers (P6−P9). There is again a large exaggeration of the short-range repulsion by the force field. Modest exaggeration of the repulsion by the DFT-D method is also seen. The SCS(MI)-MP2 and DFT-SAPT methods (in particular the latter, due to the relatively small basis set used) are typically shifted toward higher stacking energies for all intermonomer distances. Interestingly, the difference between DFT-SAPT and CBS(T) widens visibly upon reducing the intermonomer separation.

Figure 8 gives the data associated with the vertical scans of the four nonparallel (tilted) dimers with steric clashes (NP1-NP4). The NP1 geometry brings the O4 region of the

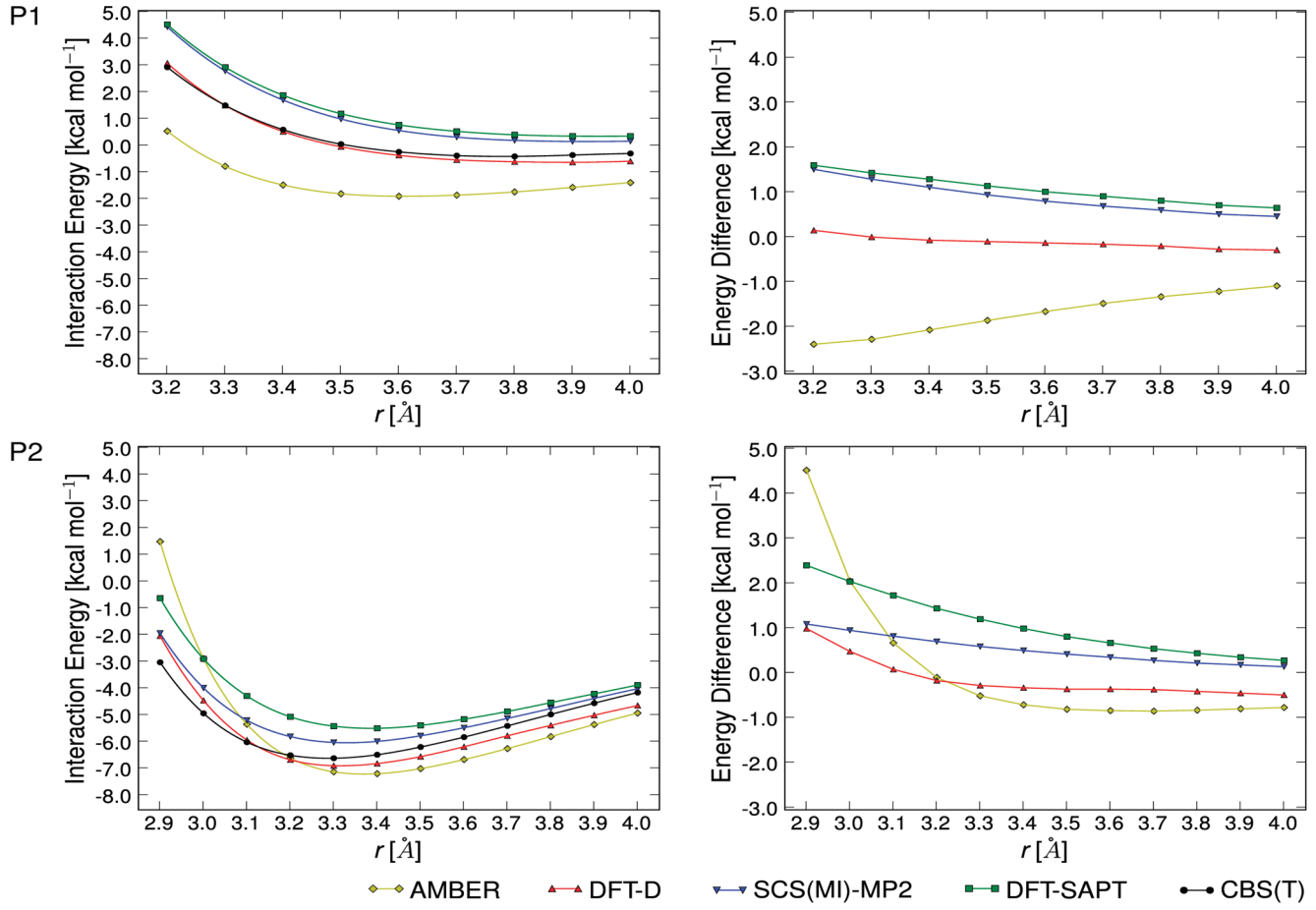


Figure 5. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for P1 and P2, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods.

first monomer into contact with the N1 region of the second one. In the NP2 geometry the N1/O2 region of the second monomer is in close contact with the aromatic ring of the first one. The NP3 geometry leads to the closest contact between the O2 atom of the second monomer and the C4 atom of the first one, and the NP4 geometry has the N3−H3 bond of the second monomer tilted to the center of the aromatic ring of the first one. The twist angle is 180° for all of the four clashed dimers. Obviously, these four dimers do not guarantee an exhaustive sampling of all possible steric contacts between the two uracils; however, we hope that they are diverse enough to allow us to make correct conclusions. The calculations are quite consistent with the results derived from the calculations on the P1−P9 dimers. The large deviation (excessive repulsion) of the AMBER force field in the short-range repulsion region is very clear, and the difference between DFT-D and CBS(T) is also visible. DFT-D overestimates the binding at larger separations, but then it has a faster and steeper onset of the repulsion upon vertical compression of the dimers. SCS(MI)-MP2 and DFT-SAPT underestimate the interaction energy and the difference between either of these methods, and CBS(T) increases as the intermolecular distance decreases. Both effects are more pronounced for DFT-SAPT, mainly because of the basis set chosen for these calculations.

Finally, Figure 9 gives the energy scans for the WC (HB1) and Calcutta (HB2) H-bonded uracil dimers. The problems of the AMBER force field to describe the repulsive region of the PES are even more visible than for the stacked dimers, in particular for the Calcutta structure. On the QM side there is essentially a very good agreement between the CBS(T), DFT-D, and SCS(MI)-MP2 methods. DFT-SAPT underestimates the strength of the H-bonding interaction at all distances. However, we need to reiterate that DFT-SAPT calculations were done with a relatively small aug-cc-pVDZ basis set. Specifically for H-bonding, very large basis sets with higher-angular momentum functions are vital.[119] The repulsion appears to be well captured by DFT-SAPT. We would like to point out again that the H-bonding data presented here is still preliminary. The issue of the description of the short-range repulsion in H-bonding is even more complex than it is for stacking and will be addressed in the future as it is beyond the scope of the present stacking study.

Interaction-Energy Statistics. rms errors with respect to the CBS(T) reference data are given in Table 1 along with the corresponding MAX values. For the twist-angle scan (P3) the best performer is the DFT-D method, with a rms error of only 0.15 kcal mol^{-1} and a MAX value of 0.23 kcal mol^{-1}. Both SCS(MI)-MP2 and DFT-SAPT underestimate the strength of the interaction across the entire range of twist angles, showing rms errors of 0.74 and 1.39 kcal mol^{-1} respectively. On the contrary, AMBER overestimates the attraction with a rms error of 0.61 kcal mol^{-1}.

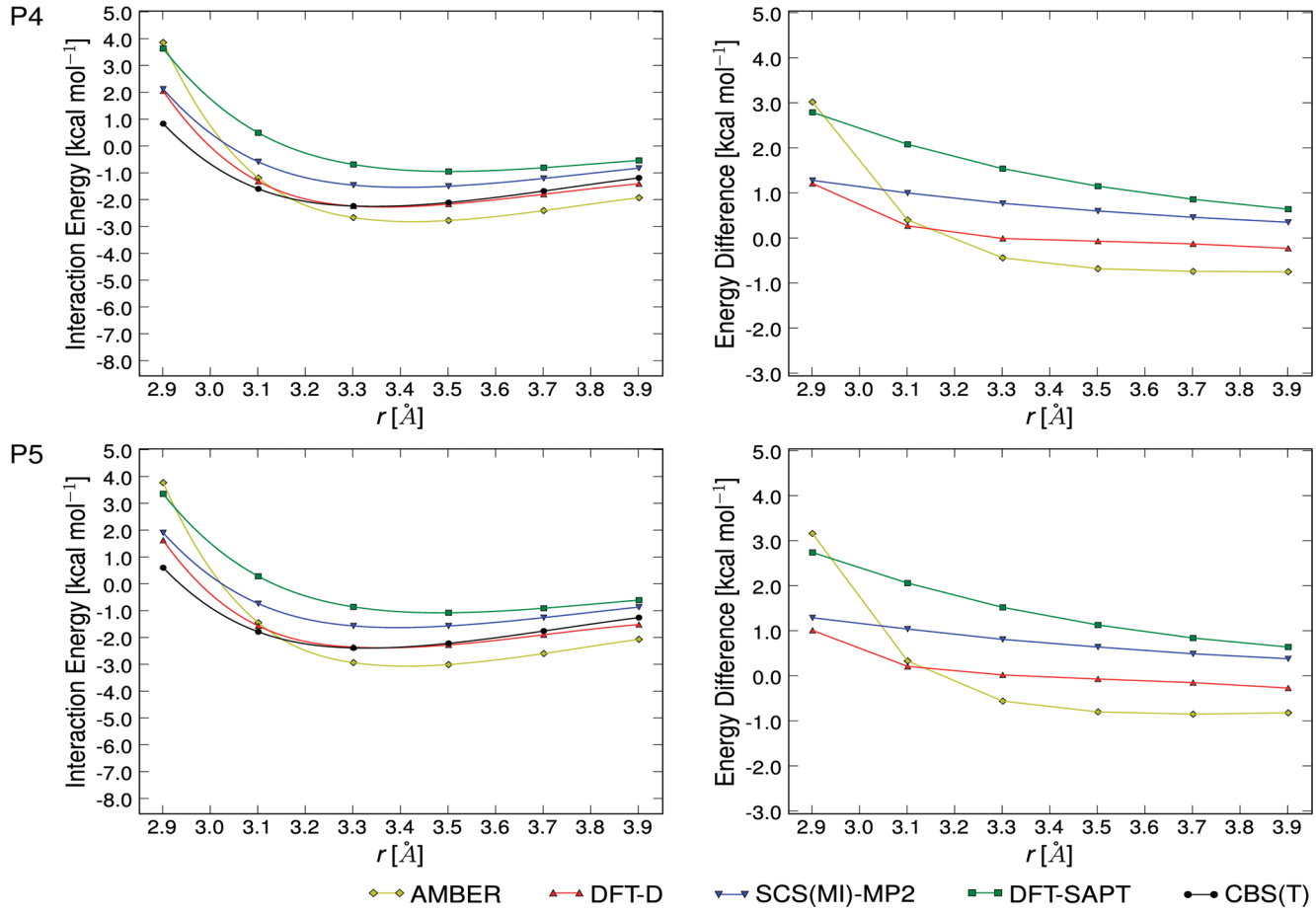


Figure 6. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for P4 and P5, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods.

Turning now to the performance of the QM methods in the vertical scans of the parallel dimers, we see that the DFT-D method is again the best performer with regards to the chosen statistical descriptors, giving for example the smallest rms deviation (0.52 kcal mol^{-1}). DFT-D overestimates the strength of the interaction at long distances and underestimates it at short distances (refer to Figures 5−8). The DFT-D deviations around the PEC minimum are typically very small. This is not surprising as the parametrization of DFT-D[53] relies mostly on a training set containing noncovalent complexes optimized to a minimum. The fact that DFT-D always underestimates the strength of the interaction at short distances may be rooted in the inability of the density functional to completely account for short-range correlation effects, which are not compensated for by the empirical correction. SCS(MI)-MP2 also shows a small rms deviation with respect to the CBS(T) data (0.86 kcal mol^{-1}), but, unlike DFT-D, the shape of the SCS(MI)-MP2 curve closely resembles the CBS(T) one for all of the scans. This means that in general this method may produce better energy gradients, in particular in the repulsive regions of the PES. This method, however, systematically underestimates the strength of the interaction. Finally, DFT-SAPT exhibits the largest rms error among the QM methods (1.37 kcal mol^{-1}), and the method systematically underestimates the strength of the interaction. Note, however, that the DFT-SAPT data were obtained with a relatively small basis set (aug-cc-pVDZ), while the other methods are calculated (directly or effectively) with a CBS extrapolation.

The AMBER force field shows larger deviations with respect to CBS(T) than the QM methods, with a rms error of 1.53 kcal mol^{-1}. AMBER overestimates the strength of the interaction at long distances and underestimates it at short distances, the only exception being the untwisted undisplaced P1 dimer, where the AMBER energies are lower along the entire range of distances. It partially could be due to increased polarization effects in this particular geometry that would not be captured by a nonpolarizable force field. However, the DFT-SAPT data below do not indicate a significant role of induction. Most likely, the overestimation of the strength of the interaction is due to the underestimation of the repulsion by AMBER, which in turn might be a consequence of an improper description of the anisotropy caused by a particularly unfavorable Pauli repulsion in the P1 conformation. The force field assumes isotropic interactions and spherical atoms, which might not be accurate enough in some specific geometries.[95]

Similar findings are seen in the case of the vertical scans for the nonparallel dimers. With a rms error of 0.55 kcal mol^{-1} DFT-D is again the best performer with respect to this statistical descriptor. The method remains to overestimate the strength of the interaction at long distances and underestimate it at short distances. The performance of SCS(MI)-MP2 is also good, bearing the smallest MAX value and a

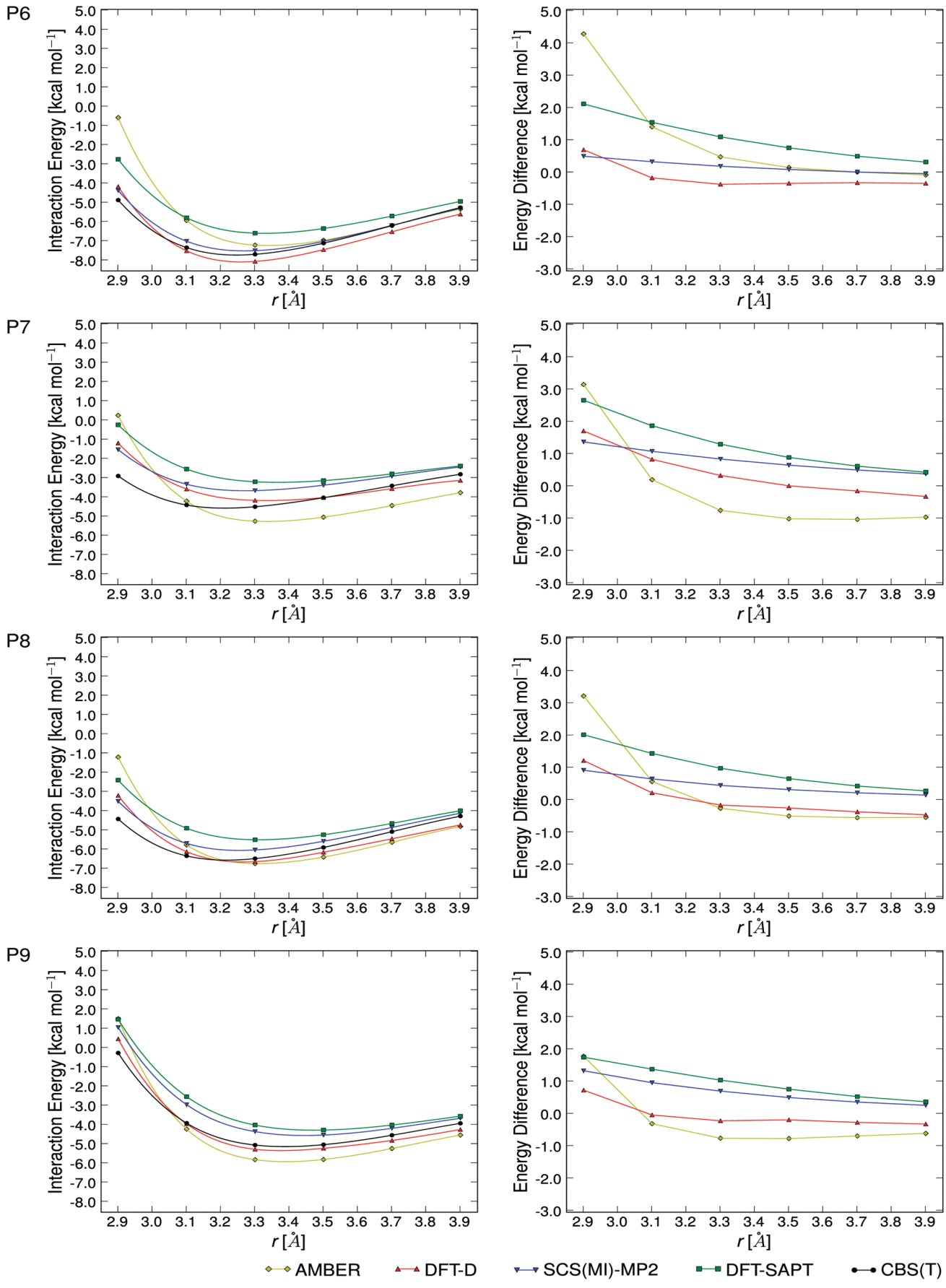


Figure 7. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for P6 through P9, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods.

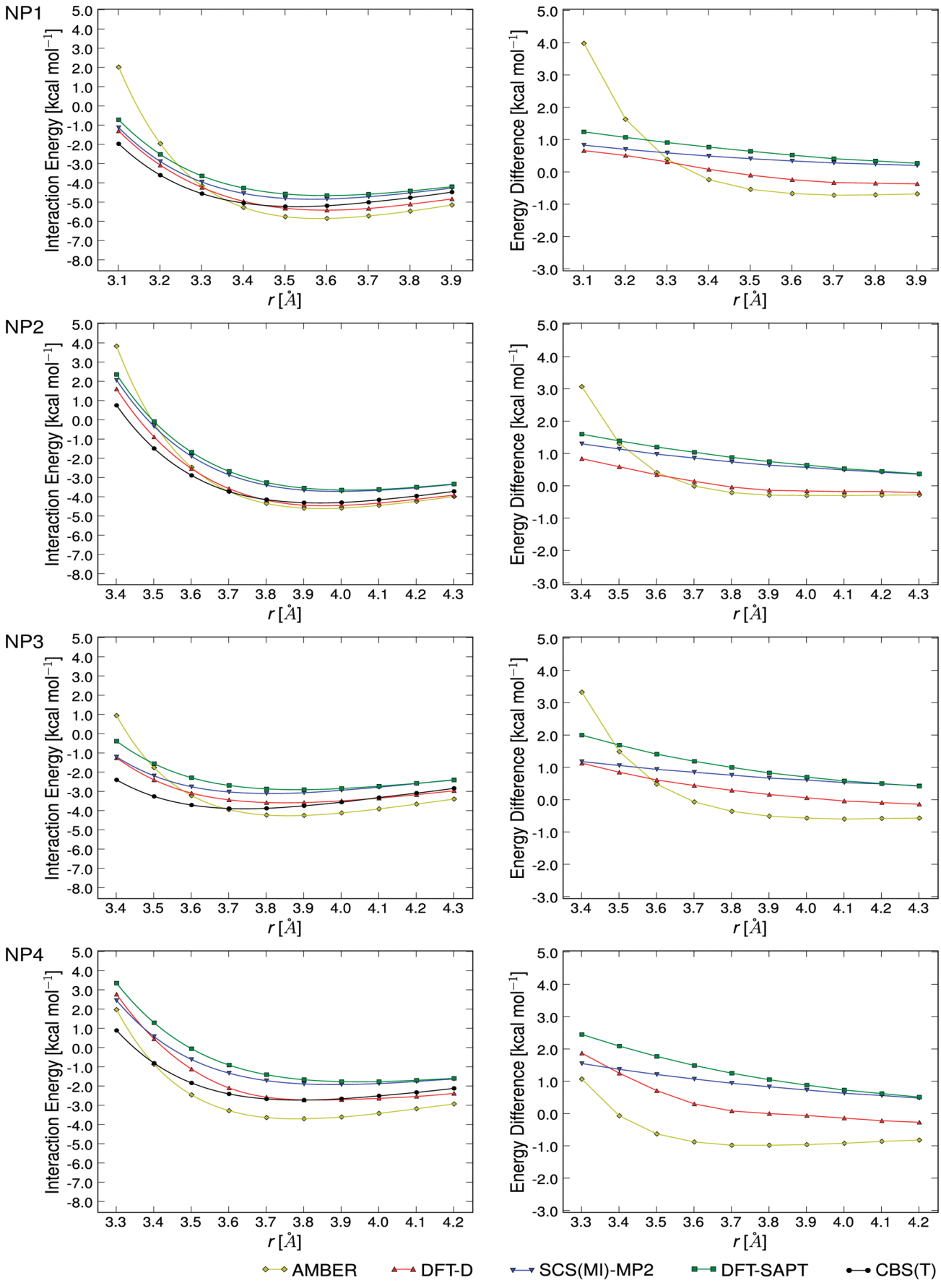


Figure 8. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for NP1 through NP4, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods.

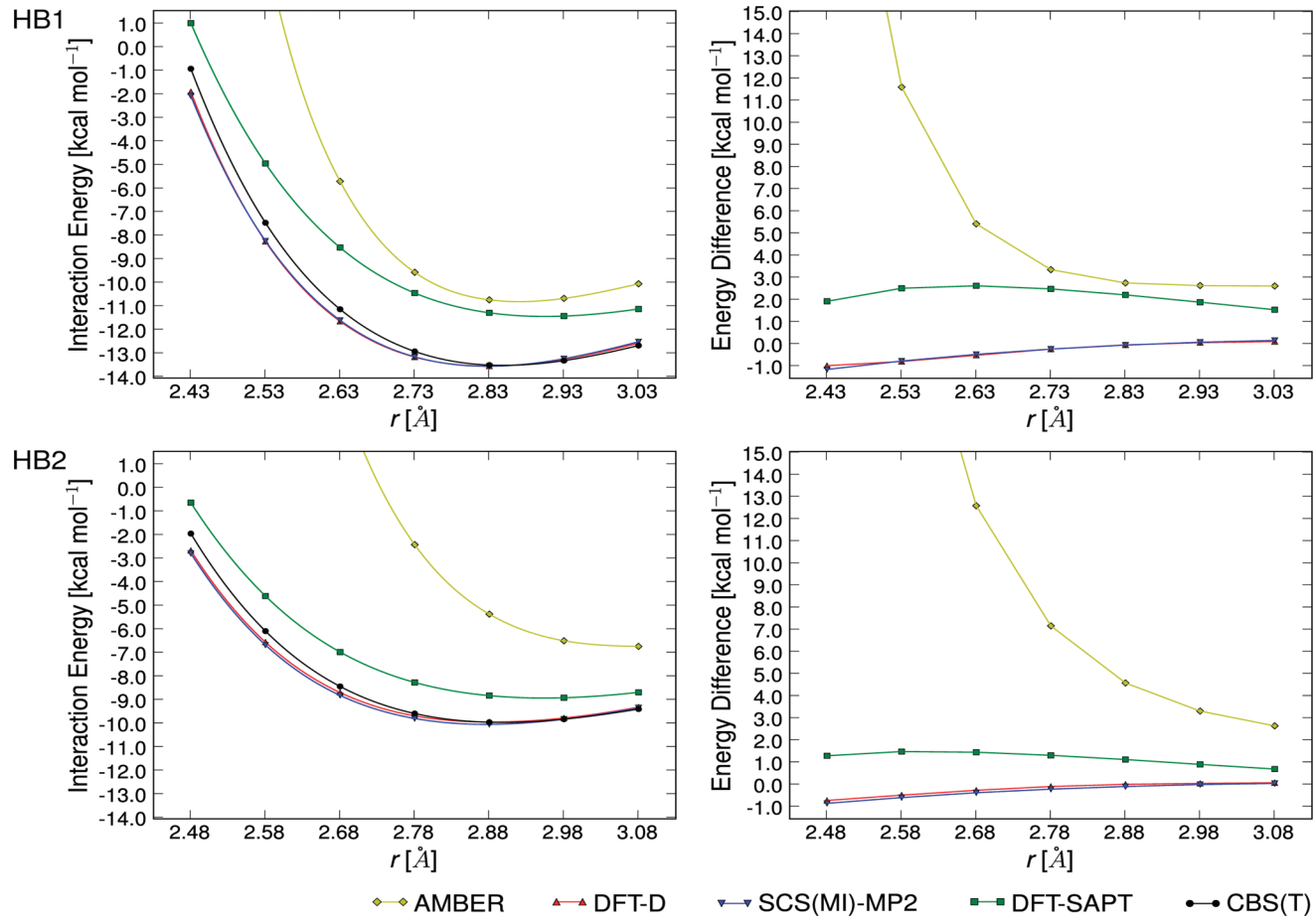


Figure 9. Potential energy curves (left) and energy differences with respect to the CBS(T) data (right) for HB1 and HB2, calculated with the AMBER, DFT-D, SCS(MI)-MP2, DFT-SAPT, and CBS(T) methods.

Table 1. Root-Mean-Squared Deviations (rmsd, kcal mol^{-1}) and Maximum Absolute Deviations (MAX, kcal mol^{-1}) of the AMBER, DFT-D, SCS(MI)-MP2, and DFT-SAPT Interaction Energies with Respect to the CBS(T) Reference Data[a]

	AMBER		DFT-D		SCS(MI)-MP2		DFT-SAPT	
dimer	rmsd	MAX	rmsd	MAX	rmsd	MAX	rmsd	MAX
P1	1.74	2.39	0.53	1.50	1.34	2.45	1.36	2.08
P2	1.58	4.52	0.45	0.99	0.60	1.09	1.27	2.40
P3	0.61	1.05	0.15	0.23	0.74	0.93	1.39	1.51
P4	1.36	3.03	0.52	1.22	0.82	1.29	1.69	2.80
P5	1.44	3.17	0.44	1.02	0.84	1.30	1.67	2.75
P6	1.85	4.29	0.41	0.70	0.26	0.50	1.22	2.12
P7	1.50	3.15	0.80	1.71	0.87	1.37	1.51	2.66
P8	1.39	3.22	0.57	1.22	0.52	0.92	1.14	2.02
P9	0.94	1.78	0.36	0.73	0.78	1.33	1.09	1.75
all P[b]	1.53	4.52	0.52	1.71	0.86	2.45	1.37	2.80
NP1	3.03	8.40	0.44	0.83	0.58	1.01	0.86	1.44
NP2	1.09	3.08	0.37	0.85	0.82	1.31	0.98	1.61
NP3	1.24	3.34	0.53	1.14	0.80	1.19	1.16	2.01
NP4	0.85	1.08	0.77	1.88	1.01	1.56	1.43	2.46
all NP	1.77	8.40	0.55	1.88	0.81	1.56	1.13	2.46
HB1	12.09	28.73	0.52	0.98	0.57	1.15	2.21	2.63
HB2	21.91	50.20	0.34	0.72	0.42	0.85	1.22	1.49
all HB	17.70	50.20	0.44	0.98	0.50	1.15	1.78	2.63

[a] Some extreme MAX values are not visualized in the figures because they are out of the range of intermolecular distances in those figures. Refer to Tables S2, S3, and S4 in the Supporting Information for the complete data. [b] P3 was not included in the calculation of rmsd and MAX values, i.e., the values in this row correspond to the group of dimers for which vertical scans have been performed.

rms error of 0.81 kcal mol^{-1}. The same error amounts to 1.13 kcal mol^{-1} in the case of DFT-SAPT. Moreover, the shapes of the SCS(MI)-MP2 curves again resemble the CBS(T) ones, and, particularly when going from the minimum to the repulsive region (see the NP4 curve in Figure 8), this method produces better energy gradients than DFT-D. AMBER shows a larger rms error of 1.77 kcal mol^{-1}, and its deviations are more pronounced in the repulsive region of the PECs. In order to assess the importance of the results, we need to consider which geometries can be

populated in real structures. It is clear that severely compressed structures are not likely to be populated, but structures with shortened intermonomer separations with energies 1−2 kcal mol^{-1} above the minima on the CBS(T) curves are certainly accessible. For these geometries the force field repulsion is already visibly exaggerated.

With respect to the H-bonding scans the best performance is achieved by the DFT-D and SCS(MI)-MP2 methods, with rmsd values of about ∼0.5 kcal mol^{-1}. Both methods overestimate the strength of the interaction at short distances. DFT-SAPT evaluated with the medium-sized basis set underestimates it at all distances, showing a rmsd value of 1.78 kcal mol^{-1}. The large rmsd associated with the AMBER data is due to the huge deviations toward more repulsive values that occur in the regions of short H-bonding distances, in particular for the HB2 structure. The very large repulsion given by the force field for this dimer indicates an excessive repulsion for the C−H···O contact. This can be due to an excessive steepness of the repulsion with the Lennard-Jones 6−12 potential and most likely is also due to a too large atomic radius of that particular hydrogen atom in the AMBER force field. The data show that the force-field performance is better for the stacked dimers than for the H-bonded ones. However, as noted in the introduction, the overestimation of short-range repulsion for H-bonded base pairs may have less serious consequences for the molecular-mechanical studies than errors in the short-range repulsion for stacking. The interaction energy of HB1 at the minimum of the AMBER curve amounts to −10.7 kcal mol^{-1}, whereas for the same geometry Šponer et al. reported an AMBER energy of −12.1 kcal mol^{-1}.[84] This difference is because the latter value was computed with the original AMBER HF charges, which differ from the MP2 charges used in the present study. The HF charge distribution is more polar, which produces a better stabilization.

Interaction-Energy Gradients. Regions of interatomic contacts are associated with interaction energy gradients, and their inaccurate description may affect the outcome of the calculations, including the population of conformations in MD simulations. Nevertheless, simulations sample also all the additional degrees of freedom which should, in most cases, attenuate the impact of the incorrect description of energy gradients in the clashed regions. In contrast, potential energy scans that were often used to study the local conformational variations can be severely distorted by the incorrect description of the repulsion as in such scans one does not vary enough degrees of freedom to overcome the exaggerated repulsion. It is thus promising to see that the QM methods perform well for the clashes. For instance, for P8 the difference between the interaction energies at r = 2.9 and r = 3.1 Å (a range that is within 0.3 Å from the CBS(T) minimum) is 1.92, 2.19, 2.50, and 2.92 kcal mol^{-1}, calculated with CBS(T), SCS(MI)-MP2, DFT-SAPT, and DFT-D, respectively. On the contrary, AMBER yields an interaction-energy difference of 4.57 kcal mol^{-1}. Clearly, the SCS(MI)-MP2 value is the closest to the CBS(T) one, and this level of accuracy might be necessary to fully understand the intrinsic interactions in nucleic-acid systems such as those mentioned above, where variations in the geometry are often accompanied by very subtle energy differences.

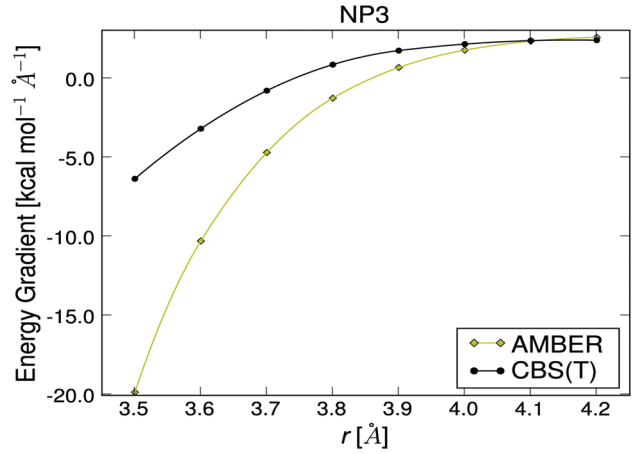


Figure 10. Estimate of the interaction-energy gradient for NP3, calculated with the AMBER and CBS(T) methods.

Figure 10 presents an estimate of the interaction-energy gradient for the AMBER and CBS(T) PECs of the NP3 dimer. The force field clearly deviates from the CBS(T) data. The NP3 geometry with vertical distance of 3.5 Å is easily accessible on the CBS(T) PEC. At this distance, however, the force field energy gradient is ∼3 times larger than the CBS(T) one, so this geometry would be penalized when attempting a force field calculation. The NP3 dimer shows probably the most severe differences among the stacked structures; similar plots for the rest of the stacked dimers can be found in the Supporting Information.

The H-bonding dimers show even more drastic differences between the AMBER and CBS(T) gradients (see Figures S12 and S13), with the same consequences as those described above for the stacked dimers.

DFT-SAPT Interaction-Energy Decomposition. Figures 11 and S14−S23 in the Supporting Information show the DFT-SAPT decompositions for the stacked dimers. Some general trends have emerged from the decomposition analysis. In basically all the stacked structures the dispersion energy is the leading stabilizing contribution, with the electrostatic term being always less stabilizing or even repulsive, as in the case of the P1 dimer at large intermolecular distances. Interestingly, for P1 the electrostatics switches from repulsion to attraction at short distances. This can be easily explained by a competition of the multipolar and overlap electrostatic components at short intermonomer distances. The multipolar part is repulsive here due to the very unfavorable dipole orientation. At short distances, however, the attractive overlap electrostatic interaction prevails. The DFT-SAPT electrostatics thus cannot be directly compared with the force field electrostatics. The latter takes into consideration only the electrostatic potentials (ESP), while the overlap electrostatic effects are effectively included in the vdW term of the force field. Therefore, the utilization of DFT-SAPT calculations in parametrizations of simple biomolecular force fields does not appear to be straightforward. It is to be noted that for biomolecular recognition the ESP part of the electrostatics is the most important term. Note that the overlap electrostatics is always

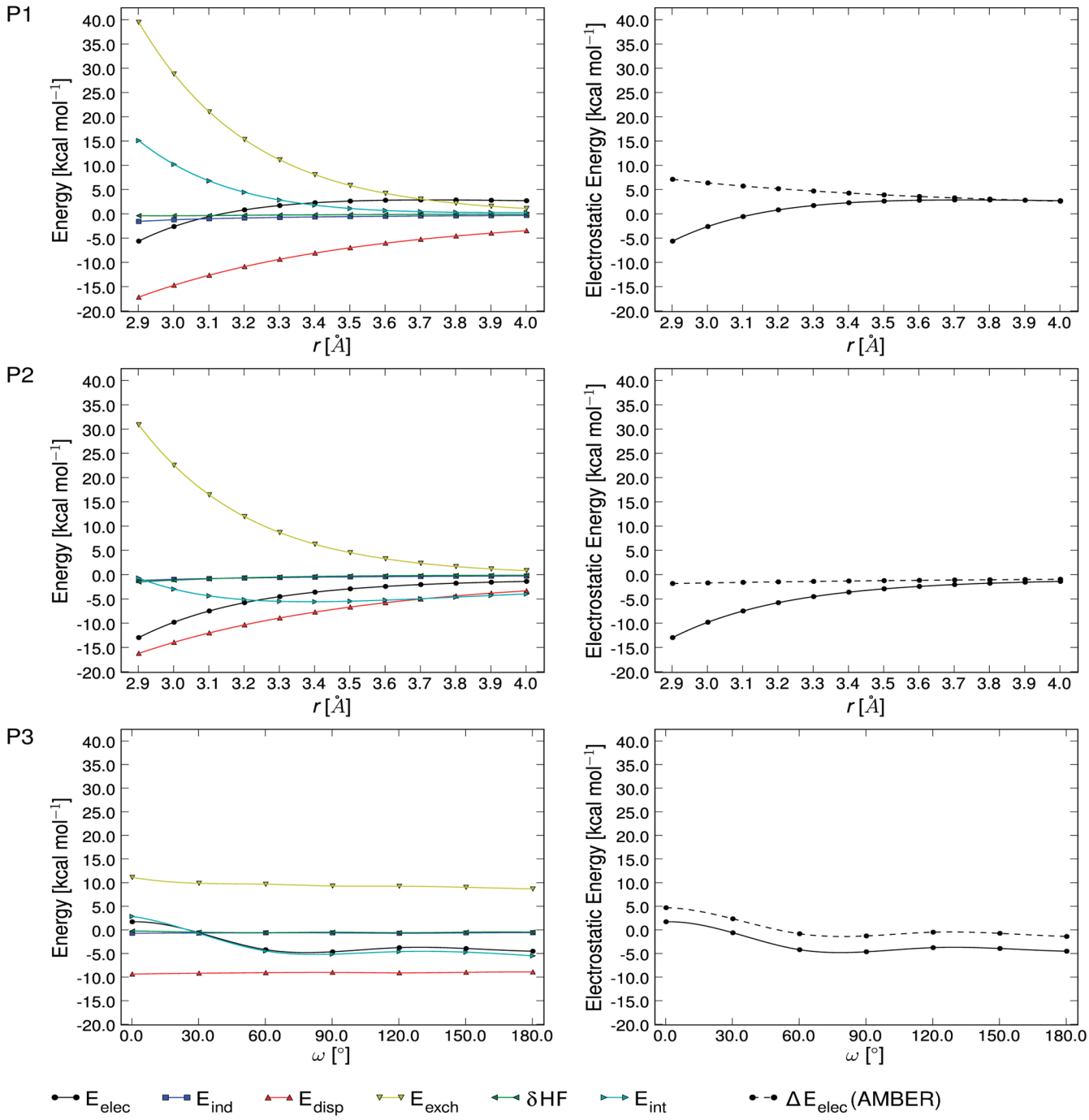


Figure 11. DFT-SAPT energy components (left) and comparison between DFT-SAPT and AMBER electrostatic energies (right) for P1, P2, and P3.

attractive, and the repulsion in the intermolecular complexes is due to Pauli exchange.

For P6, NP1, and NP2 the electrostatic interaction contribution is stabilizing and very significant, with a magnitude that is closer to that of the dispersion term, when compared to the rest of the dimers. For each dimer there is a region where the dispersion and exchange terms tend to cancel each other, usually around the minimum, which is characteristic for vdW complexes. When going into the repulsive part of the PECs the magnitude of the exchange term quickly outweighs that of the dispersion term. It thus should be noted that when assessing the DFT-SAPT decompositions, it is very important to have the dimers at proper vertical separation. This is straightforward for geometries with parallel bases but may be more complex for geometries with tilted bases. DFT-SAPT decompositions for geometries with unrelaxed vertical separation or with unrealistic clash (e.g., due to errors in the experimental structure) can be biased. Taking the variability due to the intermolecular separation into account our results are consistent with those from previous DFT-SAPT studies.[48,120,121]

For most of the stacked dimers the magnitude of the *δHF* term is comparable to that of the induction term, and both terms represent the smallest contributions to the interaction energy. The horizontal twist scan (P3) simply confirms that the electrostatic term exhibits the largest twist dependence

(followed by the exchange term to a lesser extent) and that the dispersion term is basically independent of the twist angle. Also, the dispersion and exchange terms basically cancel each other in the range of twist angles from 90 to 180°.

The DFT-SAPT data of the HB1 and HB2 structures basically confirm that the most relevant stabilizing contribution in the case of H-bonding is the electrostatic term and that the induction and δHF terms play a more important role than they do for the stacked dimers. Although smaller than the electrostatic term, the dispersion contribution is not negligible, as it amounts to about -8 and -6 kcal mol^{-1} at the CBS(T) minimum of the HB1 and HB2 dimers, respectively.

Summary and Conclusions

We have carried out reference CBS(T) calculations for about 100 geometries of the uracil dimer in stacked conformations. The calculations have been specifically aimed at geometries with unoptimized distances between the monomers including geometries with mutually tilted monomers that are characterized by a delicate balance between local steric clashes and local unstacking. Until now such geometries had not been investigated using reference QM methods and were not included in the parametrization of methods such as DFT-D. Such geometries, however, often occur in nucleic acids and are of decisive importance for local conformational variations in B-DNA as well as in some cases for the exact positioning of bases in compactly folded RNAs. Errors in the short-range repulsion region would have a major impact on potential energy scans which were often used in the past to investigate local geometry variations in DNA[64−69,76] The incorrect description of such geometries may also partially affect MD simulations in applications when quantitative accuracy is required. For comparison, we have also carried out similar "compression-extension" calculations for two H-bonded systems: the WC and the Calcutta dimers.

The results show that methods like SCS(MI)-MP2 and DFT-SAPT yield a good description of the repulsion when steric clashes are involved in the stacking of uracil dimers. These methods are thus promising to accurately scan the potential energy surfaces of these systems. However, QM studies of DNA local conformational variations are still not straightforward because the electrostatic interactions in nucleic acids are greatly reduced by the solvent screening. Most likely, calculations completely neglecting the electrostatics (equivalent to using a molecular-mechanics force field without the electrostatic term) would give a more relevant description of local B-DNA conformational variations than gas-phase calculations with full inclusion of electrostatics.[122] This is demonstrated for example in recent QM calculations of dependence of stacking on helical twist, where the optimal values of helical twist are typically outside the experimental range.[123] Both above mentioned methods also perform well for the two H-bonded dimers considered here, although for HB1 DFT-SAPT underestimates the strength of the interaction more than it does in the case of the stacked structures. Such underestimation is likely to be rooted in the relatively small basis set used for the DFT-SAPT calculations.

Given the good performance of the aforementioned methods it is natural to ask whether these methods can be further improved. For SCS(MI)-MP2 the use of the cc-pV(TQ)Z extrapolation (cc-pVTZ→cc-pVQZ) might bring a better agreement with the CBS(T) data, but there also seems to be some room for improvement on the formalism side, as it has been proposed in the work of Distasio and Head-Gordon.[77] Given the MP2 description of bond energies and intermolecular-interaction energies, it is necessary to distinguish between the computation of the two quantities, and these authors have suggested that a distance-dependent scaling could be applied to bridge these two regimes. It remains to be seen if such an approach could also result in an improved description of interaction energies of noncovalent systems in the more repulsive regions of the PES. It should be pointed out that the performance of DFT-SAPT for the systems studied here could be somewhat improved with the use of the nonlocalized PBE0AC/ALDA model as described in ref 124. The performance of DFT-SAPT would also improve with the use of better basis sets as in the present work we have used only the aug-cc-pVDZ one, which is inferior to the CBS extrapolation used for the MP2 method. In general, however, all tested QM methods have a quite reasonable performance and seem to be sufficiently accurate for the computation of intermolecular interactions in uracil dimers. Note that modest systematic shifts of the total energies in the range of ∼1 kcal/mol are less important than an eventual imbalance in the description of steric clashes and unstacking.

The DFT-D method, at least the one we tested,[53] shows somewhat larger deviations in the repulsive region. The DFT-D literature is nowadays very wide, and we do not claim that our calculations represent the full spectrum of the methods.[101,125−134] These deviations may have a modest impact on the computations, and it therefore would pay for to include some of the clashed geometries into future parametrizations of the DFT-D methods. However, we do not consider the differences as dramatic. Another approach could be a reparametrization of the empirical part of the method to improve the behavior mainly in the short-range repulsion region. However, this may require the introduction of new parameters, and success is not a priory guaranteed. The performance of DFT-D for the two tested H-bonded base pairs was neat.

We need to underline that although we expect that the present conclusions are basically valid also for other base stacking systems, it is not a priory guaranteed that the excellent performance observed here for several methods will fully transfer to all types of stacking interactions in nucleic-acid biopolymers. Thus, further calculations would be still useful.

Finally, the AMBER force field, which currently dominates molecular modeling of nucleic acids, shows large deviations in the repulsive region both for the stacked and for the H-bonded structures. The steric clashes are excessively severe, and the onset of the short-range repulsion upon compressing the interacting systems is too fast. This is expected to have a substantial effect on potential energy scans. On the other hand, we do not expect that this error would have dramatic qualitative impacts on explicit solvent simulations of nucleic acids. The reason is that in the simulations all the other degrees of freedom are also sampled, and these can be efficiently used to dissipate

the excessive clash with not much effect on the geometry. It has been noticed that an excessive repulsion associated with too large hydrogens of the thymine methyl group affects substantially 2D stacking energy scans of propeller twisting in ApT and ApA B-DNA steps. However, the reduction of the hydrogen radius, which markedly improved the quality of the empirical force-field stacking energy scans, had no visible effect in test simulations.[135] An exaggerated short-range repulsion in the force field description was also reported for cation-solute interactions and affects for example simulations of quadruplex DNA where the cations residing in the channel of the quadruplex stem look oversized. This effect was originally attributed to the lack of polarization in the force field,[136,137] but it is more likely caused by the short-range repulsion imbalance discussed in the present paper. Also, when using MP2/aug-cc-pVDZ ESP charges the AMBER force field seems to provide a better gas-phase description for stacking than for H-bonding, at least for the uracil dimer. Despite the reported difference we would like to clearly state that the overall description of base stacking and H-bonding by the nonbonded potential of the AMBER force field is good, and, as we pointed out elsewhere, stacking is probably the best approximated term in the force field.[138,139] The discrepancy reported here should not affect the overall stability of the simulations and qualitative applications of the method. However, it may affect the description of very subtle quantitative effects such as the local conformational variations in B-DNA.

Acknowledgment. This contribution was supported by the Grant Agency of the Academy of Sciences of the Czech Republic, grant nos. IAA400040802 and IAA400550701; the Ministry of Education of the Czech Republic, grant nos. LC06030, LC512, and MSM6198959216 (Petr Jurečka); the Grant Agency of the Czech Republic, 203/09/1476; and by the Academy of Sciences of the Czech Republic, grant nos. AV0Z50040507, AV0Z50040702, and Z40550506. A portion of the research described in this paper was carried out on the high-performance computer system of the Environmental Molecular Sciences Laboratory (EMSL), a scientific user facility at the Pacific Northwest National Laboratory. Pavel Hobza acknowledges the support of Praemium Academiae, Academy of Sciences of the Czech Republic.

Supporting Information Available: Studied geometries (xyz coordinates), tables of interaction energies, plots of energy gradients, and plots of DFT-SAPT energy contributions. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ng, H.; Kopka, M. L.; Dickerson, R. E. The Structure of a Stable Intermediate in the A↔B DNA Helix Transition. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 2035–2039.

(2) Mathews, D. H.; Turner, D. H. Prediction of RNA Secondary Structure by Free Energy Minimization. *Curr. Opin. Struct. Biol.* **2006**, *16*, 270–278.

(3) Petersheim, M.; Turner, D. H. Base-Stacking and Base-Pairing Contributions to Helix Stability: Thermodynamics of Double-Helix Formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry* **1983**, *22*, 256–263.

(4) Guckian, K. M.; Schweitzer, B. A.; Ren, R. X.; Sheils, C. J.; Tahmassebi, D. C.; Kool, E. T. Factors Contributing to Aromatic Stacking in Water: Evaluation in the Context of DNA. *J. Am. Chem. Soc.* **2000**, *122*, 2213–2222.

(5) Suzuki, M.; Amano, N.; Kakinuma, J.; Tateno, M. Use of a 3D Structure Data Base for Understanding Sequence-Dependent Conformational Aspects of DNA. *J. Mol. Biol.* **1997**, *274*, 421–435.

(6) Dickerson, R. E.; Drew, H. R. Structure of a B-DNA Dodecamer: II. Influence of Base Sequence on Helix Structure. *J. Mol. Biol.* **1981**, *149*, 761–786.

(7) Alexandrov, B. S.; Gelev, V.; Monisova, Y.; Alexandrov, L. B.; Bishop, A. R.; Rasmussen, K. Ø.; Usheva, A. A Nonlinear Dynamic Model of DNA with a Sequence-dependent Stacking Term. *Nucleic Acids Res.* **2009**, advance access. (Doi: 10.1093/nar/gkp016).

(8) Copeland, K. L.; Anderson, J. A.; Farley, A. R.; Cox, J. R.; Tschumper, G. S. Probing Phenylalanine/Adenine p-Stacking Interactions in Protein Complexes with Explicitly Correlated and CCSD(T) Computations. *J. Phys. Chem. B* **2008**, *112*, 14291–14295.

(9) Rutledge, L. R.; Wetmore, S. D. Remarkably Strong T-Shaped Interactions Between Aromatic Amino Acids and Adenine: Their Increase Upon Nucleobase Methylation and a Comparison to Stacking. *J. Chem. Theory Comput.* **2008**, *4*, 1768–1780.

(10) Cysewski, P. The Post-SCF Quantum Chemistry Characteristics of the Energetic Heterogeneity of Stacked Guanine-Guanine Pairs Found in B-DNA and A-DNA Crystals. *J. Mol. Struct. (THEOCHEM)* **2008**, *865*, 36–43.

(11) Lait, L. A.; Rutledge, L. R.; Millen, A. L.; Wetmore, S. D. yDNA Versus xDNA Pyrimidine Nucleobases: Computational Evidence for Dependence of Duplex Stability on Spacer Location. *J. Phys. Chem. B* **2008**, *112*, 12526–12536.

(12) Cooper, V. R.; Thonhauser, T.; Langreth, D. C. An Application of the Van Der Waals Density Functional: Hydrogen Bonding and Stacking Interactions Between Nucleobases. *J. Chem. Phys.* **2008**, *128*, 204102–4.

(13) Hill, J. G.; Platts, J. A. Calculating Stacking Interactions in Nucleic Acid Base-Pair Steps Using Spin-Component Scaling and Local Second Order Moller-Plesset Perturbation Theory. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2785–2791.

(14) Rutledge, L. R.; Durst, H. F.; Wetmore, S. D. Computational Comparison of the Stacking Interactions Between the Aromatic Amino Acids and the Natural or (Cationic) Methylated Nucleobases. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2801–2812.

(15) Langner, K. M.; Sokalski, W. A.; Leszczynski, J. Intriguing Relations of Interaction Energy Components in Stacked Nucleic Acids. *J. Chem. Phys.* **2007**, *127*, 111102–4.

(16) Vanommeslaeghe, K.; Mignon, P.; Loverix, S.; Tourwe, D.; Geerlings, P. Influence of Stacking on the Hydrogen Bond Donating Potential of Nucleic Bases. *J. Chem. Theory Comput.* **2006**, *2*, 1444–1452.

(17) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. Hybrid Density Functional Theory for p-Stacking Interactions: Application to Benzenes, Pyridines, and DNA Bases. *J. Comput. Chem.* **2006**, *27*, 491–504.

(18) Cysewski, P.; Czyznikowska-Balcerak, Z. The MP2 Quantum Chemistry Study on the Local Minima of Guanine Stacked with All Four Nucleic Acid Bases in Conformations Cor-

responding to Mean B-DNA. *J. Mol. Struct. (THEOCHEM)* **2005**, *757*, 29–36.

(19) Sponer, J.; Riley, K. E.; Hobza, P. Nature and Magnitude of Aromatic Stacking of Nucleic Acid Bases. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595–2610.

(20) Sponer, J.; Leszczynski, J.; Hobza, P. Electronic Properties, Hydrogen Bonding, Stacking, and Cation Binding of DNA and RNA Bases. *Biopolymers* **2001**, *61*, 3–31.

(21) Sponer, J.; Jurecka, P.; Hobza, P. Base Stacking and Base Pairing. In *Computational studies of RNA and DNA*; Sponer, J., Lankas, F., Eds.; Springer: Dordrecht, 2006; Chapter *14*, pp 343−388.

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

(23) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. Quadratic Configuration Interaction. A General Technique for Determining Electron Correlation Energies. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(24) Paldus, J.; Shavitt, I.; Cizek, J. Correlation Problems in Atomic and Molecular Systems. IV. Extended Coupled-Pair Many-Electron Theory and Its Application To the BH_3 Molecule. *Phys. Rev. A* **1972**, *5*, 50–67.

(25) DeJong, E. S.; Marzluff, W. F.; Nikonowicz, E. P. NMR Structure and Dynamics of the RNA-Binding Site for the Histone mRNA Stem-Loop Binding Protein. *RNA* **2002**, *8*, 83–96.

(26) Theimer, C. A.; Finger, L. D.; Trantirek, L.; Feigon, J. Mutations Linked to Dyskeratosis Congenita Cause Changes in the Structural Equilibrium in Telomerase RNA. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 449–454.

(27) Kolk, M. H.; Graaf, M. V. D.; Wijmenga, S. S.; Pleij, C. W. A.; Heus, H. A.; Hilbers, C. W. NMR Structure of a Classical Pseudoknot: Interplay of Single- and Double-Stranded RNA. *Science* **1998**, *280*, 434–438.

(28) Theimer, C. A.; Blois, C. A.; Feigon, J. Structure of the Human Telomerase RNA Pseudoknot Reveals Conserved Tertiary Interactions Essential for Function. *Mol. Cell* **2005**, *17*, 671–682.

(29) Liu, H.; Matsugami, A.; Katahira, M.; Uesugi, S. A Dimeric RNA Quadruplex Architecture Comprised of Two G:G(:A):G:G(:A) Hexads, G:G:G:G Tetrads and UUUU Loops. *J. Mol. Biol.* **2002**, *322*, 955–970.

(30) Oberstrass, F. C.; Lee, A.; Stefl, R.; Janis, M.; Chanfreau, G.; Allain, F. H. Shape-Specific Recognition in the Structure of the Vts1p SAM Domain with RNA. *Nat. Struct. Mol. Biol.* **2006**, *13*, 160–167.

(31) Sashital, D. G.; Venditti, V.; Angers, C. G.; Cornilescu, G.; Butcher, S. E. Structure and Thermodynamics of a Conserved U2 snRNA Domain from Yeast and Human. *RNA* **2007**, *13*, 328–338.

(32) Larson, S. B.; Day, J.; Greenwood, A.; McPherson, A. Refined Structure of Satellite Tobacco Mosaic Virus at 1.8 Å Resolution. *J. Mol. Biol.* **1998**, *277*, 37–59.

(33) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* **2000**, *289*, 905–20.

(34) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. Structure of the 70S Ribosome Complexed with mRNA and tRNA. *Science* **2006**, *313*, 1935–42.

(35) Sponer, J.; Leszczynski, J.; Hobza, P. Nature of Nucleic Acid-Base Stacking: Nonempirical *Ab Initio* and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. *J. Phys. Chem.* **1996**, *100*, 5590–5596.

(36) Hehre, W. J.; Radom, L.; Schleyer, P. V. R.; Pople, J. A. In *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; pp 1−548.

(37) Hobza, P.; Sponer, J.; Polasek, M. H-Bonded and Stacked DNA Base Pairs: Cytosine Dimer. An *Ab Initio* Second-Order Moeller-Plesset Study. *J. Am. Chem. Soc.* **1995**, *117*, 792–798.

(38) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(39) Kratochvil, M.; Engkvist, O.; Sponer, J.; Jungwirth, P.; Hobza, P. Uracil Dimer: Potential Energy and Free Energy Surfaces. *Ab Initio* Beyond Hartree-Fock and Empirical Potential Studies. *J. Phys. Chem. A* **1998**, *102*, 6921–6926.

(40) Kratochvil, M.; Engkvist, O.; Vacek, J.; Jungwirth, P.; Hobza, P. Methylated Uracil Dimers: Potential Energy and Free Energy Surfaces. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2419–2424.

(41) Hobza, P.; Sponer, J. Significant Structural Deformation of Nucleic Acid Bases in Stacked Base Pairs: An *Ab Initio* Study Beyond Hartree-Fock. *Chem. Phys. Lett.* **1998**, *288*, 7–14.

(42) Florian, J.; Sponer, J.; Warshel, A. Thermodynamic Parameters for Stacking and Hydrogen Bonding of Nucleic Acid Bases in Aqueous Solution: *Ab Initio*/Langevin Dipoles Study. *J. Phys. Chem. B* **1999**, *103*, 884–892.

(43) Leininger, M. L.; Nielsen, I. M. B.; Colvin, M. E.; Janssen, C. L. Accurate Structures and Binding Energies for Stacked Uracil Dimers. *J. Phys. Chem. A* **2002**, *106*, 3850–3854.

(44) Hobza, P.; Sponer, J. Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

(45) Czyznikowska, Z.; Zalesny, R.; Ziolkowski, M.; Gora, R. W.; Cysewski, P. The Nature of Interactions in Uracil Dimer: An *Ab Initio* Study. *Chem. Phys. Lett.* **2007**, *450*, 132–137.

(46) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(47) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(48) Cybulski, H.; Sadlej, J. Symmetry-Adapted Perturbation-Theory Interaction-Energy Decomposition for Hydrogen-Bonded and Stacking Structures. *J. Chem. Theory Comput.* **2008**, *4*, 892–897.

(49) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of Van Der Waals Complexes. *Chem. Rev.* **1994**, *94*, 1887–1930.

(50) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. Intermolecular Potentials Based on Symmetry-Adapted Perturbation Theory with Dispersion Energies from Time-

Dependent Density-Functional Calculations. *J. Chem. Phys.* **2005**, *123*, 214103–14.

(51) Pitonak, M.; Riley, K. E.; Neogrady, P.; Hobza, P. Highly Accurate CCSD(T) and DFT-SAPT Stabilization Energies of H-Bonded and Stacked Structures of the Uracil Dimer. *ChemPhysChem* **2008**, *9*, 1636–1644.

(52) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(53) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. Density Functional Theory Augmented with an Empirical Dispersion Term. Interaction Energies and Geometries of 80 Noncovalent Complexes Compared with *Ab Initio* Quantum Mechanics Calculations. *J. Comput. Chem.* **2007**, *28*, 555–569.

(54) Yanson, I. K.; Teplitsky, A. B.; Sukhodub, L. F. Experimental Studies of Molecular Interactions Between Nitrogen Bases of Nucleic Acids. *Biopolymers* **1979**, *18*, 1149–1170.

(55) Jurecka, P.; Hobza, P. True Stabilization Energies for the Optimal Planar Hydrogen-Bonded and Stacked Structures of Guanine. .Cytosine, Adenine. .Thymine, and Their 9- and 1-Methyl Derivatives: Complete Basis Set Calculations at the MP2 and CCSD(T) Levels and Comparison with Experiment. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

(56) Casaes, R. N.; Paul, J. B.; McLaughlin, R. P.; Saykally, R. J.; van Mourik, T. Infrared Cavity Ringdown Spectroscopy of Jet-Cooled Nucleotide Base Clusters and Water Complexes. *J. Phys. Chem. A* **2004**, *108*, 10989–10996.

(57) Ts'o, P. O. P.; Melvin, I. S.; Olson, A. C. Interaction and Association of Bases and Nucleosides in Aqueous Solutions. *J. Am. Chem. Soc.* **1963**, *85*, 1289–1296.

(58) Calladine, C. R. Mechanics of Sequence-Dependent Stacking of Bases in B-DNA. *J. Mol. Biol.* **1982**, *161*, 343–352.

(59) Calladine, C. R.; Drew, H. R. A Base-Centred Explanation of the B-to-A Transition in DNA. *J. Mol. Biol.* **1984**, *178*, 773–782.

(60) Yanagi, K.; Privé, G. G.; Dickerson, R. E. Analysis of Local Helix Geometry in Three B-DNA Decamers and Eight Dodecamers. *J. Mol. Biol.* **1991**, *217*, 201–214.

(61) Dickerson, R. E.; Goodsell, D. S.; Neidle, S. ". .the tyranny of the lattice. .". *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 3579–3583.

(62) Ferre-D'Amare, A. R.; Zhou, K. H.; Doudna, J. A. Crystal Structure of a Hepatitis Delta Virus Ribozyme. *Nature (London)* **1998**, *395*, 567–574.

(63) Krasovska, M. V.; Sefcikova, J.; Reblova, K.; Schneider, B.; Walter, N. G.; Sponer, J. Cations and Hydration in Catalytic RNA: Molecular Dynamics of the Hepatitis Delta Virus Ribozyme. *Biophys. J.* **2006**, *91*, 626–638.

(64) Bhattacharyya, D.; Bansal, M. Local Variability and Base Sequence Effects in DNA Crystal Structures. *J. Biomol. Struct. Dyn.* **1990**, *8*, 539–572.

(65) Bhattacharyya, D.; Bansal, M. Groove Width and Depth of B-DNA Structures Depend on Local Variation in Slide. *J. Biomol. Struct. Dyn.* **1992**, *10*, 213–226.

(66) Sarai, A.; Mazur, J.; Nussinov, R.; Jernigan, R. L. Origin of DNA Helical Structure and Its Sequence Dependence. *Biochemistry* **1988**, *27*, 8498–8502.

(67) Srinivasan, A. R.; Torres, R.; Clark, W.; Olson, W. K. Base Sequence Effects in Double Helical DNA. I. Potential Energy Estimates of Local Base Morphology. *J. Biomol. Struct. Dyn.* **1987**, *5*, 459–96.

(68) Sponer, J.; Kypr, J. Different Intrastrand and Interstrand Contributions to Stacking Account for Roll Variations at the Alternating Purine-Pyrimidine Sequences in A-DNA and A-RNA. *J. Mol. Biol.* **1991**, *221*, 761–764.

(69) Sponer, J.; Kypr, J. Relationships Among Rise, Cup, Roll and Stagger in DNA Suggested by Empirical Potential Studies of Base Stacking. *J. Biomol. Struct. Dyn.* **1993**, *11*, 27–41.

(70) Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C.; Westhof, E.; Wolberger, C.; Berman, H. M. A Standard Reference Frame for the Description of Nucleic Acid Base-Pair Geometry. *J. Mol. Biol.* **2001**, *313*, 229–237.

(71) Reblova, K.; Lankas, F.; Razga, F.; Krasovska, M. V.; Koca, J.; Sponer, J. Structure, Dynamics and Elasticity of free 16S rRNA Helix 44 Studied by Molecular Dynamics Simulations. *Biopolymers* **2006**, *82*, 504–520.

(72) Lankas, F.; Sponer, J.; Langowski, J.; Cheatham, T. E. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.* **2003**, *85*, 2872–2883.

(73) Dock-Bregeon, A. C.; Chevrier, B.; Podjarny, A.; Johnson, J.; de Bear, J. S.; Gough, G. R.; Gilham, P. T.; Moras, D. Crystallographic Structure of an RNA Helix: $[U(UA)_6A]_2$. *J. Mol. Biol.* **1989**, *209*, 459–474.

(74) Wahl, M. C.; Sundaralingam, M. Crystal Structures of A-DNA Duplexes. *Biopolymers* **1997**, *44*, 45–63.

(75) Sponer, J.; Jurecka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps. *Chem.--Eur. J.* **2006**, *12*, 2854–2865.

(76) Sponer, J.; Kypr, J. Theoretical Analysis of the Base Stacking in DNA: Choice of the Force Field and a Comparison with the Oligonucleotide Crystal Structures. *J. Biomol. Struct. Dyn.* **1993**, *11*, 277–292.

(77) Distasio, R. A.; Head-Gordon, M. Optimized Spin-Component Scaled Second-Order Møller-Plesset Perturbation Theory for Intermolecular Interaction Energies. *Mol. Phys.* **2007**, *105*, 1073–1083.

(78) Hesselmann, A.; Jansen, G. The Helium Dimer Potential from a Combined Density Functional Theory and Symmetry-Adapted Perturbation Theory Approach Using an Exact Exchange-Correlation Potential. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010–5014.

(79) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. Assessment of the Performance of the M05−2X and M06−2X Exchange-Correlation Functionals for Noncovalent Interactions in Biomolecules. *J. Chem. Theory Comput.* **2008**, *4*, 1996–2000.

(80) Bachorz, R. A.; Bischoff, F. A.; Höfener, S.; Klopper, W.; Ottiger, P.; Leist, R.; Frey, J. A.; Leutwyler, S. Scope and Limitations of the SCS-MP2 Method for Stacking and Hydrogen Bonding Interactions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2758–2766.

(81) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of Approximate Integrals in *Ab Initio* Theory. An Application in MP2 Energy Calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(82) Vahtras, O.; Almlof, J.; Feyereisen, M. W. Integral Approximations for LCAO-SCF Calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(83) Bernholdt, D. E.; Harrison, R. J. Large-Scale Correlated Electronic Structure Calculations: The RI-MP2 Method on Parallel Computers. *Chem. Phys. Lett.* **1996**, *250*, 477–484.

(84) Sponer, J.; Jurecka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

(85) Van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. State of the Art in Counterpoise Theory. *Chem. Rev.* **1994**, *94*, 1873–1885.

(86) Szalewicz, K.; Jeziorski, B. Comment on "On the Importance of the Fragment Relaxation Energy Terms in the Estimation of the Basis Set Superposition Error Correction to the Intermolecular Interaction Energy" [J. Chem. Phys. 104, 8821 (1996)]. *J. Chem. Phys.* **1998**, *109*, 1198–1200.

(87) Jansen, H. B.; Ros, P. Non-Empirical Molecular Orbital Calculations on the Protonation of Carbon Monoxide. *Chem. Phys. Lett.* **1969**, *3*, 140–143.

(88) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(89) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129–145.

(90) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431–439.

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

(92) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(93) Reha, D.; Kabelac, M.; Ryjacek, F.; Sponer, J.; Sponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. Intercalators. 1. Nature of Stacking Interactions Between Intercalators (Ethidium, Daunomycin, Ellipticine, and 4′, 6-Diaminide-2-phenylindole) and DNA Base Pairs. *Ab initio* Quantum Chemical, Density Functional Theory, and Empirical Potential Study. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.

(94) Jurecka, P.; Sponer, J.; Hobza, P. Potential Energy Surface of the Cytosine Dimer: MP2 Complete Basis Set Limit Interaction Energies, CCSD(T) Correction Term and a Comparison with the AMBER Force Field. *J. Phys. Chem. B* **2004**, *108*, 5466–5471.

(95) Sponer, J.; Leszczynski, J.; Hobza, P. Base Stacking in Cytosine Dimer. A Comparison of Correlated *Ab Initio* Calculations with Three Empirical Potential Models and Density Functional Theory Calculations. *J. Comput. Chem.* **1996**, *17*, 841–850.

(96) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic Structure Calculations on Workstation Computers: The Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(97) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(98) Baerends, E. J.; Ellis, D. E.; Ros, P. Self-Consistent Molecular Hartree--Fock--Slater Calculations I. The Computational Procedure. *Chem. Phys.* **1973**, *2*, 41–51.

(99) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(100) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. On Some Approximations in Applications of Xa Theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(101) Grimme, S. Accurate Description of Van Der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(102) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401−1146401−4.

(103) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson T. *MOLPRO, Version 2006.1, A Package of Ab Initio Programs*. See http://www.molpro.net (accessed month year).

(104) Polly, R.; Werner, H.; Manby, F. R.; Knowles, P. J. Fast Hartree-Fock Theory Using Local Density Fitting Approximations. *Mol. Phys.* **2004**, *102*, 2311–2321.

(105) Werner, H.; Manby, F. R.; Knowles, P. J. Fast Linear Scaling Second-Order Moller-Plesset Perturbation Theory (MP2) Using Local and Density Fitting Approximations. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

(106) Grimme, S. Improved Second-Order Moller-Plesset Perturbation Theory by Separate Scaling of Parallel- and Antiparallel-Spin Pair Correlation Energies. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

(107) Hesselmann, A.; Jansen, G. First-Order Intermolecular Interaction Energies from Kohn-Sham Orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

(108) Hesselmann, A.; Jansen, G. Intermolecular Induction and Exchange-Induction Energies from Coupled-Perturbed Kohn-Sham Density Functional Theory. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

(109) Hesselmann, A.; Jansen, G. Intermolecular Dispersion Energies from Time-Dependent Density Functional Theory. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

(110) Hesselmann, A.; Jansen, G.; Schutz, M. Density-Functional Theory-Symmetry-Adapted Intermolecular Perturbation Theory with Density Fitting: A New Efficient Method to Study Intermolecular Interaction Energies. *J. Chem. Phys.* **2005**, *122*, 014103–17.

(111) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(112) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-Set Convergence in Correlated Calculations on Ne, N2, and H2O. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(113) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. Basis-Set Convergence of the Energy in Molecular Hartree-Fock Calculations. *Chem. Phys. Lett.* **1999**, *302*, 437–446.

(114) Jurecka, P.; Hobza, P. On the Convergence of the (DECCS-D(T) -DEMP2) Term for Complexes with Multiple H-Bonds. *Chem. Phys. Lett.* **2002**, *365*, 89–94.

(115) Dabkowska, I.; Jurecka, P.; Hobza, P. On Geometries of Stacked and H-Bonded Nucleic Acid Base Pairs Determined at Various DFT, MP2, and CCSD(T) Levels up to the CCSD(T)/Complete Basis Set Limit Level. *J. Chem. Phys.* **2005**, *122*, 204322–9.

(116) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comp. Sci. Eng.* **2007**, *9*, 90–95.

(117) Jones, E.; Oliphant, T.; Peterson, P. and others. SciPy: Open Source Scientific Tools for Python ; 2001. http://www.scipy.org (accessed month year).

(118) Oliphant, T. E. Python for Scientific Computing. *Comp. Sci. Eng.* **2007**, *9*, 10–20.

(119) Sponer, J.; Hobza, P. Interaction Energies of Hydrogen Bonded Formamide Dimer, Formamidine Dimer, and Selected DNA Base Pairs Obtained with Large Basis Sets of Atomic Orbitals. *J. Phys. Chem. A* **2000**, *104*, 4592–4597.

(120) Hesselmann, A.; Fitzgerald, G.; Schutz, M. Interaction Energy Contributions of H-Bonded and Stacked Structures of the AT and GC DNA Base Pairs from the Combined Density Functional Theory and Intermolecular Perturbation Theory Approach. *J. Am. Chem. Soc.* **2006**, *128*, 11730–11731.

(121) Sedlak, R.; Jurecka, P.; Hobza, P. Density Functional Theory-Symmetry Adapted Perturbation Treatment Energy Decomposition of Nucleic Acid Base Pairs Taken from DNA Crystal Geometry. *J. Chem. Phys.* **2007**, *127*, 075104.

(122) Sponer, J.; Florian, J.; Ng, H.; Sponer, J. E.; Spackova, N. Local Conformational Variations Observed in B-DNA Crystals Do Not Improve Base Stacking: Computational Analysis of Base Stacking in a d(CATGGGCCCATG)2 B↔A Intermediate Crystal Structure. *Nucleic Acids Res.* **2000**, *28*, 4893–4902.

(123) Cooper, V. R.; Thonhauser, T.; Puzder, A.; Schroder, E.; Lundqvist, B. I.; Langreth, D. C. Stacking Interactions and the Twist of DNA. *J. Am. Chem. Soc.* **2008**, *130*, 1304–1308.

(124) Tekin, A.; Jansen, G. How Accurate Is the Density Functional Theory Combined with Symmetry-Adapted Perturbation Theory Approach for CH-π and π-π Interactions? A Comparison to Supermolecular Calculations for the Acetylene-Benzene Dimer. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1680–1687.

(125) Ducere, J.; Cavallo, L. Parametrization of an Empirical Correction Term to Density Functional Theory for an Accurate Description of p-Stacking Interactions in Nucleic Acids. *J. Phys. Chem. B* **2007**, *111*, 13124–13134.

(126) Goursot, A.; Mineva, T.; Kevorkyants, R.; Talbis, D. Interaction Between N-Alkane Chains: Applicability of the Empirically Corrected Density Functional Theory for Van Der Waals Complexes. *J. Chem. Theory Comput.* **2007**, *3*, 755–763.

(127) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(128) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. Dispersion Corrections to Density Functionals for Water Aromatic Interactions. *J. Chem. Phys.* **2004**, *120*, 2693–2699.

(129) Wu, Q.; Yang, W. Empirical Correction to Density Functional Theory for Van Der Waals Interactions. *J. Chem. Phys.* **2002**, *116*, 515–524.

(130) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. Towards Extending the Applicability of Density Functional Theory to Weakly Bound Systems. *J. Chem. Phys.* **2001**, *115*, 8748–8757.

(131) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. Hydrogen Bonding and Stacking Interactions of Nucleic Acid Base Pairs: A Density-Functional-Theory Based Treatment. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(132) Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Eijck, B. P. Transferable *Ab Initio* Intermolecular Potentials. 1. Derivation from Methanol Dimer and Trimer Calculations. *J. Phys. Chem. A* **1999**, *103*, 9872–9882.

(133) Meijer, E. J.; Sprik, M. A Density-Functional Study of the Intermolecular Interactions of Benzene. *J. Chem. Phys.* **1996**, *105*, 8684–8689.

(134) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duartes, H. A. An Efficient a Posteriori Treatment for Dispersion Interaction in Density-Functional-Based Tight Binding. *J. Chem. Theory Comput.* **2005**, *1*, 841–847.

(135) Warmlander, S.; Sponer, J. E.; Sponer, J.; Leijon, M. The Influence of the Thymine C5 Methyl Group on Spontaneous Base Pair Breathing in DNA. *J. Biol. Chem.* **2002**, *277*, 28491–28497.

(136) Spackova, N.; Berger, I.; Sponer, J. Nanosecond Molecular Dynamics Simulations of Parallel and Antiparallel Guanine Quadruplex DNA Molecules. *J. Am. Chem. Soc.* **1999**, *121*, 5519–5534.

(137) Fadrna, E.; Spackova, N.; Stefl, R.; Koca, J.; Cheatham, T. E.; Sponer, J. Molecular Dynamics Simulations of Guanine Quadruplex Loops: Advances and Force Field Limitations. *Biophys. J.* **2004**, *87*, 227–242.

(138) Sponer, J.; Spackova, N. Molecular Dynamics Simulations and Their Application to Four-Stranded DNA. *Methods* **2007**, *43*, 278–290.

(139) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. Molecular Dynamics Simulations of RNA: An *In Silico* Single Molecule Approach. *Biopolymers* **2007**, *85*, 169–184.

CT9000125

JCTC Journal of Chemical Theory and Computation

Ab Initio Density Fitting: Accuracy Assessment of Auxiliary Basis Sets from Cholesky Decompositions

Jonas Boström,[†] Francesco Aquilante,[‡] Thomas Bondo Pedersen,[†] and Roland Lindh*,[†]

Department of Theoretical Chemistry, Chemical Center, University of Lund, P.O. Box 124 S-221 00 Lund, Sweden, and Department of Physical Chemistry, Sciences II, University of Geneva, Quai E. Ansermet 30, 1211 Geneva 4, Switzerland

Received January 15, 2009

Abstract: The accuracy of auxiliary basis sets derived by Cholesky decompositions of the electron repulsion integrals is assessed in a series of benchmarks on total ground state energies and dipole moments of a large test set of molecules. The test set includes molecules composed of atoms from the first three rows of the periodic table as well as transition metals. The accuracy of the auxiliary basis sets are tested for the 6-31G**, correlation consistent, and atomic natural orbital basis sets at the Hartree−Fock, density functional theory, and second-order Møller−Plesset levels of theory. By decreasing the decomposition threshold, a hierarchy of auxiliary basis sets is obtained with accuracies ranging from that of standard auxiliary basis sets to that of conventional integral treatments.

1. Introduction

The density fitting (DF) or resolution-of-the-identity (RI) approximation[1] is an efficient approach for speeding up quantum chemical calculations. Gaussian auxiliary basis sets for the fitting procedure have been optimized and extensively tested with respect to the accuracy of ground state energies for Hartree−Fock (HF) theory, nonhybrid as well as hybrid density functional theory (DFT), and second-order Møller−Plesset (MP2) theory.[2−11] The goals of the optimizations were to keep errors due to the DF approximation below the inherent basis set incompleteness error for each theoretical model, while limiting the number of auxiliary functions to a few times the number of atomic orbital (AO) basis functions.

Aiming at an accurate approximation of each individual integral, we have recently proposed generating auxiliary basis sets by Cholesky decomposition (CD) of the two-electron integral matrix in AO basis. The construction of the auxiliary basis set thus becomes a purely numerical procedure carried out on-the-fly. In the Full-CD approach the entire molecular integral matrix is decomposed, and the resulting auxiliary basis set consists of both one- and two-center functions.[12,13] The one-center CD (1C-CD) approximation is obtained by restricting the decomposition of the molecular integrals such that only one-center functions enter the auxiliary basis set.[14] To reduce the computational cost of obtaining the auxiliary basis set, the atomic CD (aCD) set is obtained by a decomposition of the atomic integral matrix.[14] The atomic compact CD (acCD) auxiliary basis set is obtained from the aCD by removing linear dependence among the primitive Gaussians, again by CD.[15] We thus have available a hierarchy of *ab initio* DF approximations with an accuracy controlled by a single parameter, the CD threshold (τ). The adjective *ab initio* underlines the fact that no additional information is needed to perform a DF calculation with CD-based auxiliary basis sets compared to the corresponding conventional calculation. The CD-based auxiliary basis sets were tested for a limited number of mostly organic molecules in the papers cited above, and it is the purpose of the present work to provide a more thorough assessment of accuracy of total ground state energies and dipole moments.

Using the Coulomb metric, the DF procedure minimizes the integral diagonal error

$$\Delta_{\mu\nu,\mu\nu} = (\mu\nu|\mu\nu) - \sum_{IK} C^{I}_{\mu\nu}(I|K)C^{K}_{\mu\nu} \qquad (1)$$

to the extent possible with a given auxiliary basis set (**C** indicates the fitting coefficients and *I,K* the auxiliary functions). The CD-based auxiliary basis sets are constructed to

* Corresponding author e-mail: roland.lindh@teokem.lu.se.
[†] University of Lund.
[‡] University of Geneva.

10.1021/ct9000284 CCC: $40.75
Published on Web 05/04/2009

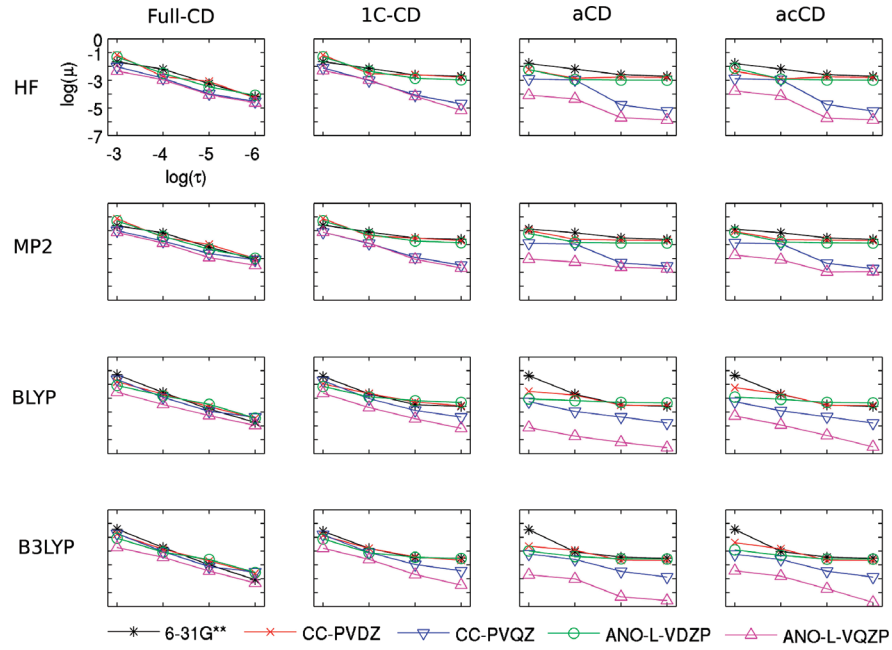


Figure 1. The mean absolute total energy errors in kcal/(mol·electron) as a function of the decomposition threshold, τ, for Set I. Each row of panels shows a specific quantum chemical method (HF, MP2, DFT/BLYP, and DFT/B3LYP), and each column of panels shows a specific CD-based auxiliary basis set (Full-CD, 1C-CD, aCD, and acCD).

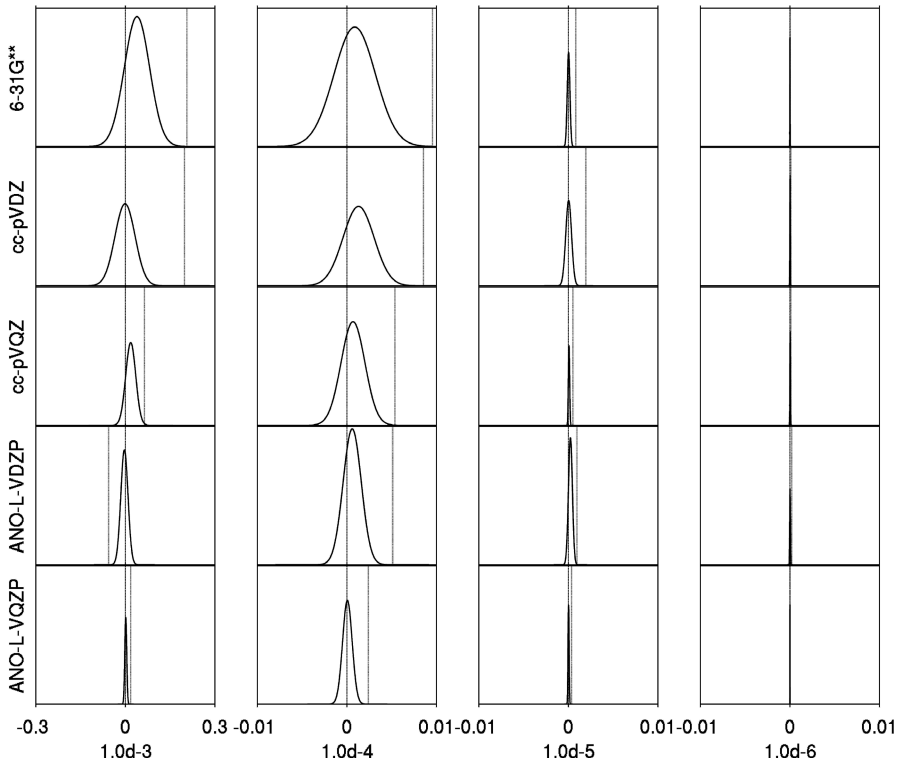


Figure 2. The mean error and standard deviation in kcal/(mol·electron) for the Full-CD DFT/B3LYP total energy calculations of Set I represented as Gaussian distributions. The dotted line represents the largest observed error. Each row of panels features a specific AO basis set, and each column of panels corresponds to a specific decomposition threshold. The scale of the ordinate is arbitrary. Note that for better visualization, the scale of the abscissa of the left most column is different than for the rest of the columns in the figure.

make the minimum value approach zero as τ is decreased. The error matrix Δ is positive semidefinite and hence

$$|\Delta_{\mu\nu,\lambda\sigma}| \leq \Delta_{\mu\nu,\mu\nu}^{1/2}\Delta_{\lambda\sigma,\lambda\sigma}^{1/2} \quad (2)$$

Full-CD guarantees that every element of the error matrix Δ is bound by τ and therefore provides complete control of the accuracy of the DF approximation.[12−14] For the one-center approximations (1C-CD, aCD, acCD), the integral

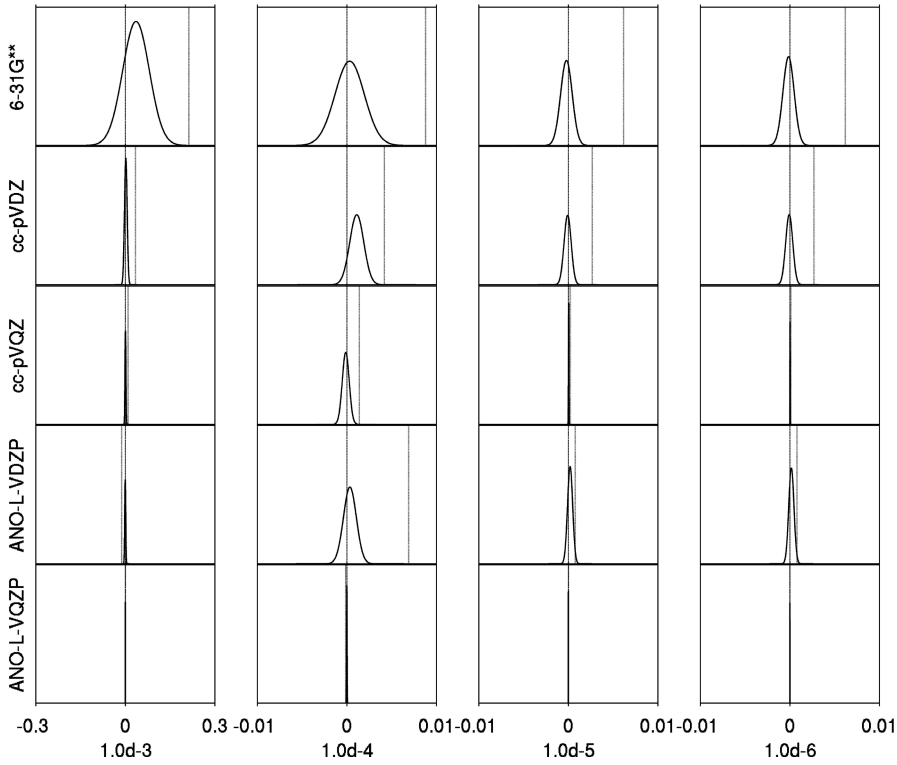


Figure 3. The mean error and standard deviation in kcal/(mol·electron) of the aCD DFT/B3LYP total energy calculations of Set I represented as Gaussian distributions. The dotted line represents the largest observed error. The scale of the ordinate is arbitrary. Each row of panels features a specific basis set, and each column of panels corresponds to a specific decomposition threshold. Note that for better visualization, the scale of the abscissa of the left most column is different than for the rest of the columns in the figure.

errors in the DF approximation to the ERIs of the type (*AA*|*AA*) and (*AA*|*BB*), where *A* and *B* are atom labels, are bound by τ, whereas integrals of the type (*AB*|**) may be affected by larger errors. The accuracy of the DF procedure with the one-center approximations is therefore limited by the ability to span these two-center functions: although the error is always minimized by the fitting procedure, it can no longer be guaranteed that the minimum value is τ.[14,15]

Here, we perform the first statistical analysis of the accuracy of the hierarchy of CD-based auxiliary basis sets: Full-CD, 1C-CD, aCD, and acCD, as a function of τ. This analysis is performed for the single configuration methods HF, pure (BLYP) and hybrid (B3LYP) DFT, and MP2 in conjunction with common segmented and generally contracted basis sets ranging from double- to quadruple-ζ levels of sophistication. We use a large test suite of molecules composed of atoms from the first three rows of the periodic table and a small set including transition metals. A similar investigation on CASSCF and CASPT2 excitation energies with CD-based auxiliary basis sets is in progress at our lab.

2. Computational Details

The purpose of this study is to test how the accuracy depends on the decomposition threshold in Full-CD,[13] 1C-CD,[14] aCD,[14] and acCD[15] for different theoretical models and AO basis sets. For the purpose of generality, the wave function models included in this study describe both Coulomb and exchange contributions and short-range correlation and dispersion. The basis set selection is representative of the wide range of AO basis sets available to computational chemists at present. Three different test sets are used in this benchmark study.

First, calculations have been performed for a large set of molecules (Set I) using four different quantum chemical methods, HF, pure and hybrid DFT, and MP2. Two functionals are used, one nonhybrid, BLYP,[16−18] and one hybrid, B3LYP.[17−19] The AO basis sets used in these calculations are Pople's 6-31G**,[20,21] Dunning's cc-pVXZ[22,23] (X = D,Q), and the ANO-L-VXZP (X = D,Q) basis sets of Widmark et al.[24,25] The values 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6} have been chosen for the Cholesky threshold. Set I is the 118 closed-shell molecules of the G2/97 test set.[26] For six of the molecules in the set we have replaced the conventional MP2 calculation with a Full-CD calculation where we have set the CD threshold to 10^{-10}.[27] For Set I, the accuracy of the HF and DFT dipole moments has also been investigated.

Second, since the molecules in Set I only include elements from the first three rows in the periodic table, a smaller set of molecules containing heavier elements is used. Specifically, Set II is composed of the seven closed-shell transition metal containing molecules of the MLBE21/05 database.[28] The accuracy assessment in association with Set II is limited to the DFT(B3LYP) model. These calculations include scalar relativistic effects through the Douglas-Kroll-Hess transformation[29−34] in conjunction with the relativistic ANO-RCC-VXZP (X = D,T) basis sets of Roos and co-workers.[35−37]

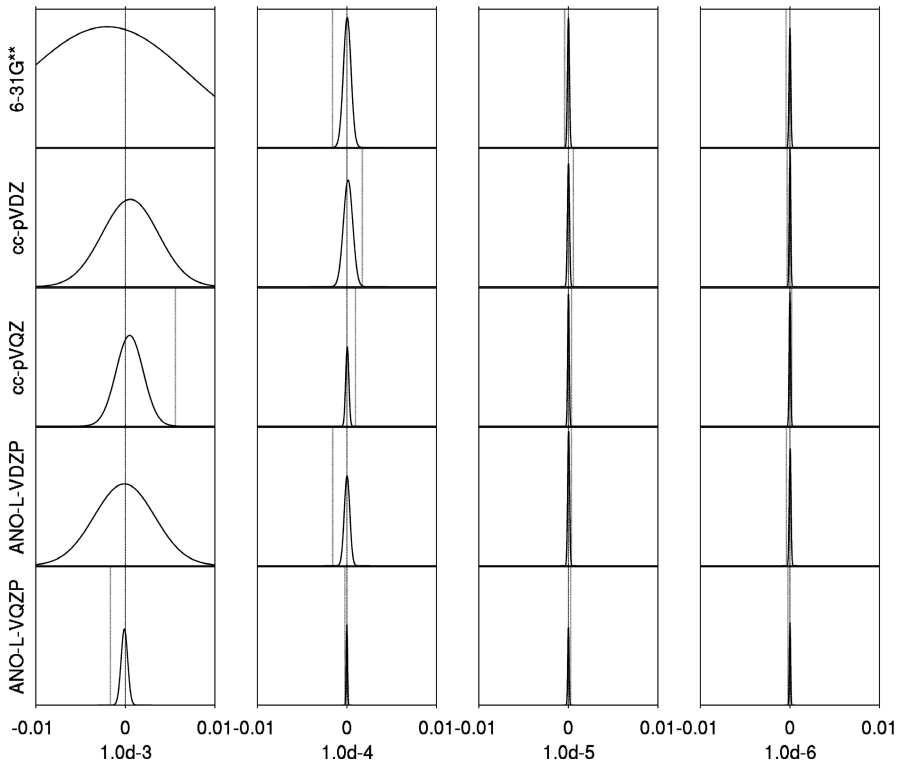


Figure 4. The mean error and standard deviation in debye of the Full-CD DFT/B3LYP dipole moment calculations of Set I represented as Gaussian distributions. The scale of the ordinate is arbitrary. The dotted line represents the largest observed error. Each row of panels features a specific AO basis set, and each column of panels corresponds to a specific decomposition threshold.

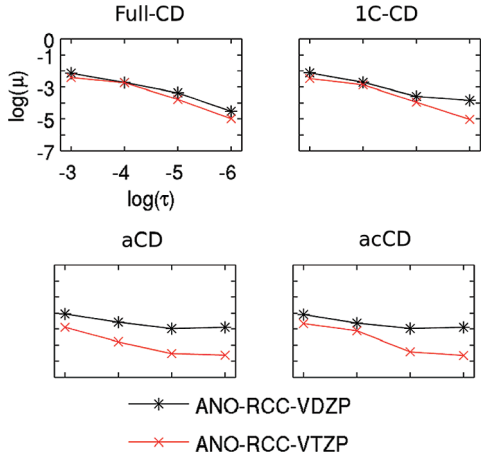


Figure 5. The mean absolute total energy errors (kcal/(mol·electron)) as a function of the value of the decomposition threshold of Set II. Each row of panels shows a specific quantum chemical method, and each column of panels shows a specific CD-based auxiliary basis set.

Third, the possibility to *ad hoc* discard the higher angular momentum auxiliary functions in association with the aCD approach is tested on a set of 24 molecules, Set III.[38] This option is considered for the 6-31G**, ANO-L-VXZP, and cc-pVXZ (X = D,T,Q) basis sets.

To measure the accuracy of the CD-based auxiliary basis sets, the error in the total energy for each molecule is computed as

$$\varepsilon_i = \frac{E_i^{conv} - E_i^{CD}}{N_i^{electrons}} \tag{3}$$

where E_i^{conv} and E_i^{CD} are the conventional and CD-based total ground state energy, respectively, and $N_i^{\mathrm{electrons}}$ is the number of electrons of molecule i. For the accuracy of the magnitude of the dipole moments, the same expression was used without a normalization against the number of electrons, as the dipole moment is a size-intensive quantity. When we present statistics on dipole moments we have also omitted every molecule with inversion symmetry since they must have zero dipole moment. The associated mean error

$$\mu_\varepsilon = \sum_i^N \frac{\varepsilon_i}{N} \tag{4}$$

mean absolute error

$$\mu_{|\varepsilon|} = \sum_i^N \frac{|\varepsilon_i|}{N} \tag{5}$$

standard deviation

$$\sigma_\varepsilon^2 = \frac{\sum_i^N \varepsilon_i - \mu_\varepsilon}{N-1} \tag{6}$$

and maximum error (with sign), $\varepsilon_{\max}$, were calculated.

All calculations have been performed using a development version of the MOLCAS quantum chemistry software.[39]

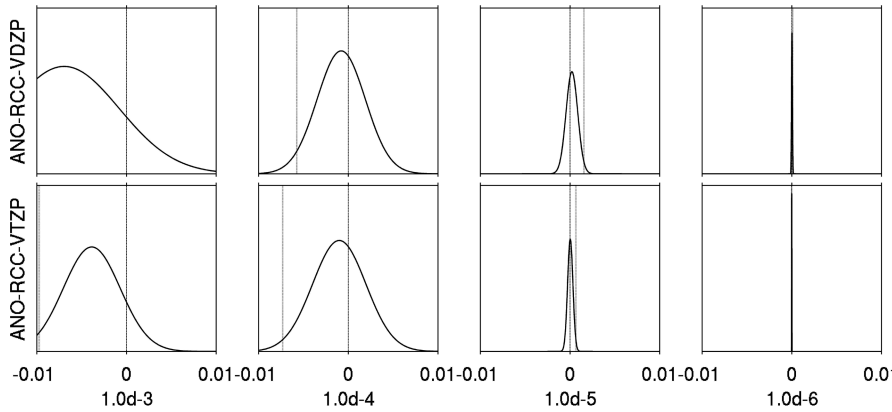


Figure 6. The mean error and standard deviation (kcal/(mol·electrons)) of the Full-CD DFT/B3LYP total energy calculations of Set II represented as Gaussian distributions. The scale of the ordinate is arbitrary. The dotted line represents the largest error in that set of molecules. Each row of panels features a specific AO basis set, and each column of panels corresponds to a specific decomposition threshold.

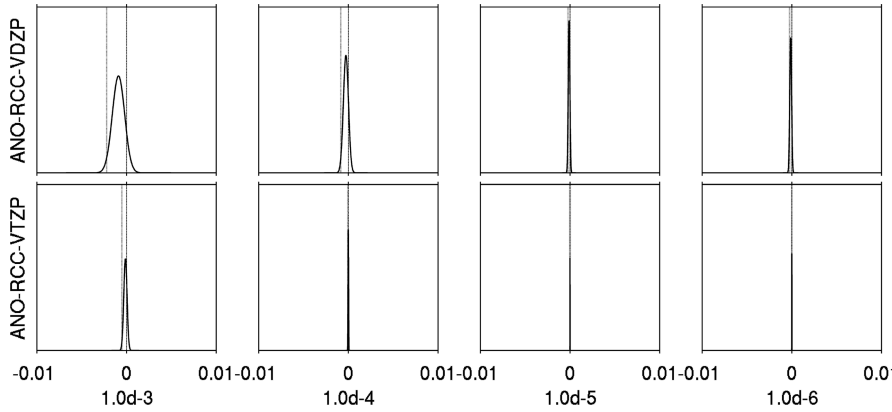


Figure 7. The mean error and standard deviation (kcal/(mol·electrons)) of the aCD DFT/B3LYP total energy calculations of Set II represented as Gaussian distributions. The scale of the ordinate is arbitrary. The dotted line represent the largest error in that set of molecules. Each row of panels features a specific AO basis set, and each column of panels a corresponds to a specific decomposition threshold.

3. Results and Discussion

In this section the results of the calculations will be visualized, and important trends and their implications for these types of calculations will be discussed. The analysis is divided into several subsections where different aspects of the results are discussed. Detailed tables of data can be found in the Supporting Information.

3.1. Independence of Theoretical Model. In Figure 1 the mean absolute errors for Set I of the studied auxiliary basis sets for various theoretical models are displayed as functions of the decomposition threshold. From Figure 1 we can conclude that the accuracy

- improves with tighter threshold,
- levels out for the one-center auxiliary basis sets (1C-CD, aCD, and acCD) for tighter thresholds, and
- is largely the same regardless of the theoretical model with a given AO basis set.

The last point confirms the assertion of our previous papers,[14,15] namely that the CD-based basis sets are unbiased. Finally, we note that the aCD and acCD accuracies are very close to each other, as expected.[15]

While the error of the Full-CD is almost linear with respect to the CD threshold (logarithmic scale), a saturation effect is observed for the one-center auxiliary basis sets. This difference is more pronounced for AO basis sets of lower quality, i.e. the double-ζ sets, for which essentially no improvement is observed for decomposition thresholds below 10^{-4}. For the quadruple-ζ basis sets, however, thresholds as low as 10^{-6} may be used with a significant gain in accuracy. The reason for this difference is that representing two-center AO products and therefore ERIs of the type $(AB|**)$ with very high accuracy places demands on the quality of the auxiliary basis set. Higher angular momentum functions are generally included in the auxiliary basis set for the quadruple-ζ AO sets than for the double-ζ ones, thus increasing accuracy of the representation of the two-center AO products. We also note that the auxiliary basis sets of the ANO-L sets tend to be more accurate than those generated from 6-31G** or Dunning's correlation consistent sets.

The observation that aCD and acCD auxiliary basis sets derived from quadruple-ζ AO basis sets are more accurate than those derived from double-ζ ones indicates that further reductions of auxiliary basis set size might be possible with a controlled loss of accuracy. This aspect will be explored below.

Table 1. Average Ratio of Auxiliary to AO Basis Functions for Each CD Method for Set I

basis set	τ	Full-CD	1C-CD	aCD/acCD
6-31G**:				
	10^{-3}	3.6	3.4	3.7
	10^{-4}	4.9	4.1	4.2
	10^{-5}	6.3	4.4	4.5
	10^{-6}	8.0	5.0	5.0
cc-pVDZ:				
	10^{-3}	3.3	3.3	4.0
	10^{-4}	4.8	4.3	4.3
	10^{-5}	6.2	4.6	4.6
	10^{-6}	7.6	5.0	5.0
cc-pVQZ:				
	10^{-3}	3.8	3.8	7.0
	10^{-4}	4.6	4.6	7.3
	10^{-5}	5.7	5.5	8.1
	10^{-6}	7.2	6.5	9.1
ANO-L-VDZP:				
	10^{-3}	3.5	3.5	5.0
	10^{-4}	4.5	4.1	5.1
	10^{-5}	5.6	4.5	5.3
	10^{-6}	7.2	5.1	5.4
ANO-L-VQZP:				
	10^{-3}	3.7	3.7	8.1
	10^{-4}	4.7	4.6	8.4
	10^{-5}	6.0	5.8	9.3
	10^{-6}	7.5	6.9	10.6

3.2. Quantitative Error Analysis. In this section a more quantitative analysis of the accuracy of the energy and the magnitude of the dipole moment for Set I will be performed. For that purpose, let the standard of Eichkorn et al.[2] be the reference. In their work they derived Coulomb fitting auxiliary basis sets associated with SVP valence basis sets (a double-ζ quality basis set) aiming at an average error of 0.2 mE_H per atom. For practical purposes this can be translated to an average error of around 0.01 kcal/(mol·electron). Moreover, an error of 0.01 debye is an acceptable level of accuracy for the computed magnitude of the dipole moment. We will in this section limit our presentation to cases which are representative and required to demonstrate the observed properties of the CD-based auxiliary basis sets. In particular, the trends of the one-center type CD auxiliary basis sets are close to identical, and only one representative, aCD, will be used to clarify the general trends and quantify average and maximum errors.

First, in Figure 2 the results from a Full-CD using DFT/B3LYP are visualized. A tighter threshold yields reduced mean and maximum errors and reduced error spread. It is clear that the three right most columns of the panels represent an average error substantially better than the norm of Eichkorn et al.[2] The left-most column (note the different scale of the abscissa) represents results on par or better than this norm. For the different AO basis sets we only note a significant difference in accuracy for the largest thresholds. From these results, we conclude that a CD threshold of 10^{-3} in association with Full-CD can be used but that a prudent user may wish to adopt a threshold one order of magnitude tighter.

Second, in Figure 3 the aCD auxiliary basis set accuracy for DFT/B3LYP is visualized. We observe that no significant improvement is achieved for thresholds below 10^{-5}, regardless of basis set. For high-quality AO basis sets, however, a substantially better accuracy is observed compared to lower-quality AO sets at the same decomposition threshold. Comparing to the Full-CD results of Figure 2, we see that the maximum error is a bit larger for aCD. We also note

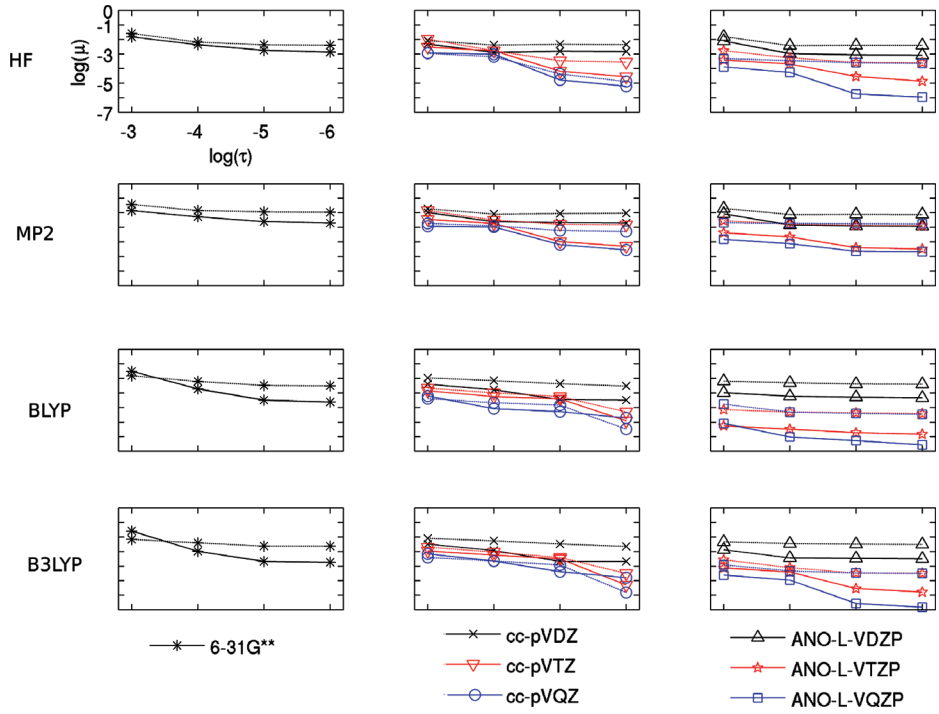


Figure 8. The aCD mean absolute energy errors (kcal/(mol·electron)) as a function of the decomposition threshold for Set III. Each row of panels shows a specific quantum chemical method. Solid lines show the results if all angular momentum components of the auxiliary basis set are included and dashed lines if some are ignored, as described in the text.

Table 2. Average Ratio of Auxiliary to AO Basis Functions for Set III When the Full Parent Product or the Reduced Product Space Is Used to Derive the aCD Based Auxiliary Basis Set

basis set	τ	full	reduced
6-31G**:			
	10^{-3}	3.7	3.0
	10^{-4}	4.1	3.2
	10^{-5}	4.3	3.5
	10^{-6}	4.8	4.0
cc-pVDZ:			
	10^{-3}	3.9	2.9
	10^{-4}	4.2	3.3
	10^{-5}	4.5	3.7
	10^{-6}	4.8	3.9
cc-pVTZ:			
	10^{-3}	5.4	3.1
	10^{-4}	5.5	3.7
	10^{-5}	6.1	4.4
	10^{-6}	6.6	5.1
cc-pVQZ:			
	10^{-3}	6.8	4.0
	10^{-4}	7.1	4.6
	10^{-5}	8.0	4.9
	10^{-6}	8.9	5.3
ANO-L-VDZP:			
	10^{-3}	4.8	3.7
	10^{-4}	4.9	3.8
	10^{-5}	5.0	3.9
	10^{-6}	5.2	4.1
ANO-L-VTZP:			
	10^{-3}	7.1	4.9
	10^{-4}	7.5	5.5
	10^{-5}	7.8	5.8
	10^{-6}	8.3	6.4
ANO-L-VQZP:			
	10^{-3}	7.8	4.9
	10^{-4}	8.1	5.2
	10^{-5}	9.1	5.7
	10^{-6}	10.1	6.6

that for the larger thresholds, the error spread is more narrow for aCD than for Full-CD.

To conclude this section, Figure 4 shows a representation of the statistics for calculations of the magnitude of the dipole moments with Full-CD at the DFT/B3LYP level of theory. Clearly, for a threshold of 10^{-4} or below an accuracy up to the standard is achieved.

3.3. Assessment for Small Transition Metal Complexes. The purpose of the calculations on Set II is to establish that the CD-based auxiliary basis sets are valid for a wider range of atoms than those represented in Set I. This section will therefore feature the same analysis as above with the purpose of demonstrating that the important features hold also for auxiliary basis sets generated from all-electron basis sets of heavy elements. Results on the magnitude of the dipole moments are collected in the Supporting Information. Figure 5 shows the mean absolute error versus the decomposition threshold.

Again, an almost linear function (logarithmic scale) is observed for Full-CD, while a saturation effect is observed for the one-center type CD-based auxiliary basis sets. The aCD and acCD accuracies are for all practical purposes identical. All of these features are the same as for the assessments of Set I, and it can be concluded that the CD-based auxiliary basis sets can be employed in DF approximations on systems with heavy elements without any loss of relative accuracy as compared to lighter elements.

From the Gaussian error distributions in Figure 6 (Full-CD, DFT/B3LYP) and Figure 7 (aCD, DFT/B3LYP) we see that the errors fall within the required norm for thresholds of 10^{-4} or tighter. In the aCD case, thresholds as high as 10^{-3} might still be practical. The results are a strong indication that the CD-based auxiliary basis sets are indeed general and can be used both for transition metals and in combination with a scalar relativistic Hamiltonian. It is reasonable to expect that this conclusion holds for any element of the periodic table.

3.4. Auxiliary Basis Set Size. It was noted above that for the highest-quality AO basis sets the DF error is smaller than one might expect from the value of the CD threshold. On the other hand, for aCDs and acCDs this can also be viewed as a downside of the approach: the aCD (or acCD) procedure generates far too many auxiliary basis functions. For efficiency purposes it is therefore mandatory to investigate the possibility of reducing the size of the aCD and acCD auxiliary basis sets.

A way to measure the efficiency of an auxiliary basis set is to look at the ratio between the number of auxiliary and AO basis functions. Eichkorn et al.[2] have in this context established that a ratio of 3 or smaller should be achievable in association with double-ζ quality basis sets. A somewhat larger ratio was established for auxiliary basis sets optimized for triple-ζ quality AO basis sets.[4] These standards are adopted here, too. However, it should be noted that while the procedure of Eichkorn et al.[2] yields an auxiliary basis set specifically designed for the DF approximation of the Coulomb potential in DFT, the aCD and acCD auxiliary basis sets are unbiased toward any quantum chemical method, as they correctly describe the AO ERIs within a certain accuracy. Hence, it should not come as a surprise that the corresponding CD-based auxiliary basis sets are somewhat larger than the standard ones. In Table 1 auxiliary to AO basis function ratios of Set I are shown for the different CD approaches.

As expected, the ratio for CD-based auxiliary basis sets with a threshold of 10^{-4} is about one-third larger than the standard achieved by Eichkorn et al. However, for the higher quality AO basis sets a substantially higher ratio is observed. These differences can be rationalized as follows. For elements of the first and second row of the periodic table, Eichkorn et al.[2] *ad hoc* eliminate the g-functions from the auxiliary basis set. This choice is based upon the knowledge of the typical structure of the density or equivalent matrices which combine with the ERIs in quantum chemical models. Obviously, the CD procedures do not have access to this information and only use the elimination of numerical linear dependence to define the size of the auxiliary basis set.

Inspired by the procedure of Eichkorn et al.,[2] we did as follows. For double-, triple-, and quadruple-ζ AO basis sets the full set of AO product functions was reduced before the aCD procedure by eliminating the g-functions, the h- and i-functions, and the i-, k-, and l-functions, respectively. The

accuracies of these reduced auxiliary basis sets as compared with auxiliary basis sets derived from the full product space of Set III are exhibited in Figure 8.

The accuracy for higher thresholds is somewhat reduced but still seems to lie within reasonable limits for a decomposition threshold of 10^{-4}, but we notice that the convergence is lost for tighter CD thresholds. Hence, the procedure should only be recommended for high threshold calculations - achieving higher accuracy is associated with a price to pay in the form of a larger auxiliary basis set. It can be seen from Table 2 that when the accuracy of the calculation permits the CD-based auxiliary basis set to be reduced, the ratios of auxiliary to AO basis functions are in parity with those of Eichkorn et al.[2] This is indeed remarkable considering that the CD-based auxiliary basis sets are nonmethod specific. However, it should still be noted that even the smallest CD-based sets, the acCD auxiliary basis sets, are larger than the standard auxiliary basis sets in terms of the number of primitive basis functions. Again, this is the price for having a nonmethod specific auxiliary basis set.

4. Summary

The accuracy of the Full-CD, 1C-CD, aCD, and acCD auxiliary basis sets has been investigated with respect to the CD threshold. The tests have been performed on a large array of molecules containing elements from the first three rows of the periodic table and transition metals. The analysis is based on the error in the total ground state energy and the magnitude of the dipole moment. It is confirmed that the CD approximations are unbiased and form a hierarchy of approximations going from an accuracy of standard auxiliary basis sets to that of a conventional two-electron integral treatment. It is also demonstrated that the CD approach to auxiliary basis set generation is most accurate with AO basis sets of high quality. The investigation shows that a CD threshold of 10^{-4} is a reasonable standard corresponding to an absolute error of less than 0.01 kcal/(mol·electron), although a CD threshold of 10^{-3} can in some cases produce results with an acceptable accuracy. The CD-based auxiliary basis sets are more computationally demanding than preoptimized sets, as they contain more auxiliary functions. This is the price to pay for an unbiased auxiliary basis set.

Acknowledgment. Funding from the Swiss National Science Foundation (SNF), the Swedish Research Council (VR), and the Linnaeus Project "Organizing Molecular Matter" at Lund University is gratefully acknowledged.

Supporting Information Available: Tables containing energy and dipole moment mean errors, mean absolute errors, standard deviation, and max errors according to eqs 4, 5, and 6 for Set I and Set II. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Vahtras, O.; Almlöf, J.; Feyereisen, M. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(2) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(3) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652–660.

(4) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(5) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

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

(7) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

(8) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.

(9) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(10) Hellweg, A.; Hättig, C.; Höfener, S.; Klopper, W. *Theor. Chem. Acc.* **2007**, *117*, 587–597.

(11) Weigend, F. *J. Comput. Chem.* **2008**, *29*, 167–175.

(12) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

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

(14) Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2007**, *127*, 114107.

(15) Aquilante, F.; Pedersen, T. B. Gagliardi, L.; Lindh, R. *J. Chem. Phys.* **2009**, *130*, 154107.

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

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

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

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

(20) Hariharan, P.; Pople, J. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(21) Francl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

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

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

(24) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.

(25) Widmark, P.-O.; Persson, B. J.; Roos, B. O. *Theor. Chim. Acta* **1991**, *79*, 419–432.

(26) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1071.

(27) Due to memory issues in the conventional MP2 calculations a Full-CD calculation with a 10^{-10} Cholesky threshold was instead used as a reference in the following molecules: two species of C_4H_{10} (trans-butane and isobutane), C_5H_8 (spiropentane), C_6H_6 (benzene), $(CH_3)_2CHOH$ (isopropanol), and $(CH_3)_3N$ (trimethylamine).

(28) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127–11143.

(29) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 2037–2047.

(30) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945–10956.

(31) Reiher, M. *Theor. Chem. Acc.* **2006**, *116*, 241–252.

(32) Wolf, A.; Reiher, M.; Hess, B. *J. Chem. Phys.* **2002**, *117*, 9215–9226.

(33) Wolf, A.; Reiher, M. *J. Chem. Phys.* **2006**, *124*, 06102.

(34) Wolf, A.; Reiher, M. *J. Chem. Phys.* **2006**, *124*, 06103.

(35) Roos, B. O.; Veryazov, V.; Widmark, P.-O. *Theor. Chem. Acc.* **2004**, *111*, 345–351.

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

(37) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

(38) Set III contains the molecules LiH, CH_2, CH_4, NH_3, H_2O, FH, SiH_2, SiH_4, PH_3, SH_2, HCl, LiF, C_2H_2, C_2H_4, C_2H_6, HCN, CO, H_2CO, H_3COH, N_2, H_2NNH_2, H_2O_2, F_2, and CO_2 from the G2/97 test suite.

(39) 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.

CT9000284

JCTC Journal of Chemical Theory and Computation

Comparison of the Properties of Bent and Straight Single-Walled Carbon Nanotube Intramolecular Junctions

Bingchun Xue, Xueguang Shao, and Wensheng Cai*

Department of Chemistry, Nankai University, Tianjin 300071, P.R. China

Received January 21, 2009

Abstract: The properties of four finite-length bent and straight intramolecular junctions (IMJs) connecting two armchair and zigzag single-walled carbon nanotube segments, viz. (3,3)-(6,0) and (4,4)-(8,0), were investigated. Their structures were calculated using the density functional theory (DFT) methods at the B3LYP/6-31G(d) level of theory. The results indicate that the bent junctions are more stable than the straight ones due to the energetically favored defect structures. Remarkable differences of the HOMO and LUMO orbitals appear between the straight and the bent IMJs. The spin-unrestricted calculations at the same level of theory were also performed to obtain the antiferromagnetic-type ground state, suggesting that the spin polarizations mainly occur on the zigzag edge and the defect rings of the straight (4,4)-(8,0) IMJ and induce marked changes of the electronic structures. Additionally, the energy band structures of the four junctions with periodic boundary conditions were calculated based on DFT calculations using generalized gradient approximation with the Perdew and Wang function. The calculated band gaps suggest that the conductance of the straight IMJs is higher than the bent ones.

Introduction

Single-wall carbon nanotubes (SWCNTs) are formed by rolling up a section of a single graphite sheet. Depending on the orientation of the roll-up vector, the SWCNTs can be metallic or semiconducting. Due to the extraordinary electronic properties, many experimental and theoretical studies have been carried out to investigate their potential use in nanoscale devices. The research has also revealed that the connections of SWCNTs with different diameter and chirality into intramolecular junctions (IMJs)[1,2] may be an important step in the development of carbon based nanoelectronic devices, because these materials are able to function as molecular diodes,[3,4] rectifiers,[5,6] electronic switches,[7,8] and so on. For the realization of nanometer-scale electronic devices, an appropriate electrical conductivity is required, which is directly related to the corresponding geometrical and electronic structures. Therefore, studies on the structures and electronic properties are important. The IMJs are generally composed of two SWCNT segments, jointed by pentagon and heptagon defects located at the interface to maintain topological consequences. The corresponding properties vary with the features of the individual segments and the amount and location of the defects at the joint part. The stabilities of the IMJs of two zigzag nanotubes and of two armchair segments have been investigated, respectively, and the most stable structures were suggested in these studies through analyzing the junctions connected by different defects.[9−11] In addition, the electronic properties of IMJs formed by different chiral tubes have also been studied theoretically,[12−15] and experimentally.[16,17] Particularly, the effect of varying the length of a metal-semiconductor IMJ on the local density of states (LDOS) was calculated.[11] In addition, for the hydrogen-terminated finite-length SWCNTs, the existence of hydrogen passivated zigzag edges has been reported to induce spin polarization on the zigzag edges.[18−21] It indicates that under the influence of an external electric field, these systems may become half-metallic with one spin channel and act as spin filters. Yet, the effect of the defects on the geometry and electronic structure and the effect of the hydrogen atoms on the spin polarization for the IMJs when joining a metallic tube to a metallic or a semiconducting tube are still fuzzy.

In this paper, the structures and band properties of the IMJs connected from armchair and zigzag nanotubes, (3,3)-(6,0)

* Corresponding author e-mail: wscai@nankai.edu.cn.

10.1021/ct900039v CCC: $40.75
Published on Web 05/07/2009

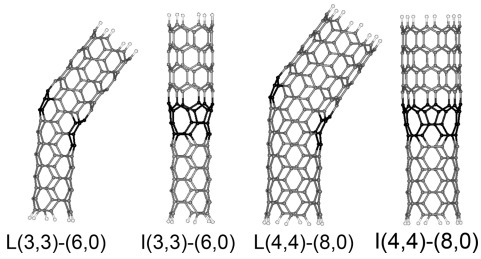


Figure 1. Optimized structures of IMJs at the B3LYP/6-31G(d) theoretical level. The pentagon and heptagon rings at the joint part are visualized in black.

and (4,4)-(8,0) were investigated, respectively. Generally, all armchair and zigzag (*n*,0) (where *n* is a multiple of 3) SWCNTs are metallic. Therefore, the IMJs discussed here are considered to be a metallic-metallic and a metallic-semiconducting junction. To examine the effect of the defects, straight- and bent-shaped structures for each junction were constructed and subsequently optimized using the B3LYP/6-31G(d) method. Their stability and frontier molecular orbitals were compared. For investigation of the possible spin polarization, the calculations with the spin-unrestricted approach were also carried out. Imposing the periodic boundary conditions to the IMJ structures, their corresponding band structures were explored by means of the generalized gradient approximation (GGA) with the Perdew and Wang function (PW91). The value of the band gap may help us to understand the conductivity of these IMJs, which is important for the future application of nanoscale electric devices.

Computational Methods

In the IMJs of (3,3)-(6,0) and (4,4)-(8,0), the corresponding two component segments in one junction have the similar diameter (4.38 ± 0.31 Å and 5.86 ± 0.44 Å). Various shaped IMJs can be formed when connecting the two segments by different position and amount of five- and seven-membered rings. In this contribution, only two typical structures, viz. bent and straight junctions, are taken into account.

The bent structures include only one pentagon and one heptagon located on the opposite position of the tube circumference. In the straight structures, the pentagons and heptagons are alternately placed on the mismatching region. According to the shape of IMJs, the bent and straight IMJs are denoted as L and I type, and four IMJs are then distinguished as L(3,3)-(6,0), I(3,3)-(6,0), L(4,4)-(8,0), and I(4,4)-(8,0), respectively, as depicted in Figure 1.

All the initially constructed structures of IMJs were optimized based on the tight binding potential for carbon. The dangling bonds at the two open ends were then saturated by hydrogen atoms in order to simulate finite nanotubes. In order to reduce the influence of the terminated hydrogen atoms on the junction region, more than three layers of carbon rings in each segment were modeled. These structures were used as starting points for the further optimization at the B3LYP/6-31G(d) theoretical level using the Gaussian 03 program packages.[22] The relative stabilities and geometric properties of IMJs were analyzed based on this level of theory.

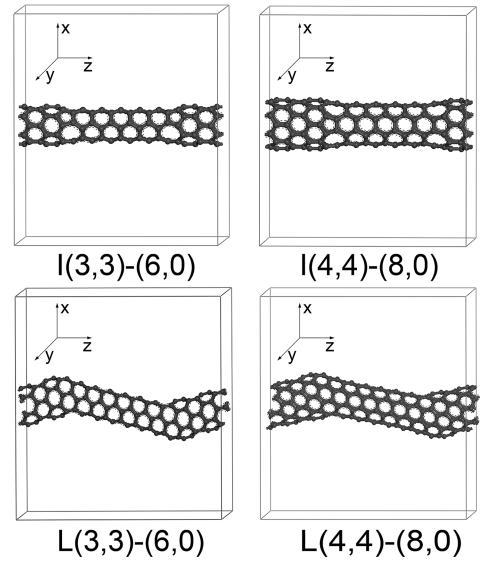


Figure 2. Schematic diagram of the unit cells used to construct the infinite-length IMJs.

Due to the previous reports that spin polarization may arise from the existence of hydrogen passivated zigzag SWCNT and local structural defects,[18−21] the unrestricted B3LYP (UB3LYP) approach with the 6-31G(d) basis set was also applied to optimize the four finite-length IMJs. Since the previous studies[20,21] show that antiferromagnetic ordering is energetically favored compared to the higher spin multiplicity state, only the antiferromagnetic-type spin is calculated in this contribution. In order to obtain the antiferromagnetic-type ground state, the symmetry of the initial guess was destroyed using the Guess=(Always, Mix) keyword in the corresponding Gaussian calculations.

Moreover, to eliminate the influence of terminated hydrogen and calculate the band structures, the infinite-length IMJs were built by periodically repeating the unit cell along the *z* axis, as seen in Figure 2. Each unit cell consisted of two junction parts separated by four carbon rings to avoid the influence between each other. In the *z* direction, the length of the cell was adjusted to make the bond lengths between the atoms located in the adjacent unit cells reasonable, so that a perfect infinite-length IMJs can be obtained. In addition, in the *x* and *y* direction, a large lattice constant (35 Å for the bent and 30 Å for the straight) was applied to prevent a significant interaction between the tubes. The optimizations of the unit cells and the calculations of the band structures were performed using the Dmol^3 package,[23,24] which has been successfully applied to study electronic properties of carbon nanotubes.[25−27] All-electron calculations were performed with the double numerical basis set and the GGA method with PW91 function. The employed convergence criteria for structural optimizations are 1×10^{-4} Hartree and 0.05 eV/Å for the energy and maximum displacement, respectively. Eleven Monkhorst-Pack k-points for the Brillouin zone integration along the *z* axis were used.

Results and Discussions

Relative Stability. The calculation results of the four IMJs in Figure 1 using the B3LYP/6-31G(d) method are given in

Table 1. Results of the Four IMJs at the B3LYP/6-31G(d) Level[a]

	molecule	energy	HOMO	LUMO	ΔE
L(3,3)-(6,0)	$C_{114}H_{12}$	−4350.549	−0.152	−0.120	0.032
I(3,3)-(6,0)	$C_{114}H_{12}$	−4350.496	−0.152	−0.133	0.019
L(4,4)-(8,0)	$C_{156}H_{16}$	−5954.204	−0.148	−0.134	0.014
	$C_{152}H_{16}$	−5801.732*			
I(4,4)-(8,0)	$C_{152}H_{16}$	−5801.670	−0.142	−0.135	0.007

[a] ΔE is the HOMO-LUMO energy gap. All the energies are in units of Hartree. The value marked with the asterisk is calculated by the energy of $C_{156}H_{16}$ subtracting the energy of four carbon atoms.

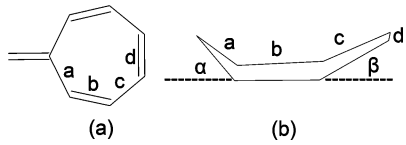


Figure 3. (a) 7-Methylene-1,3,5-cycloheptatriene used as a model molecule. (b) Schematic diagram of heptagon defects in the IMJs. The labels are used to define the bonds and dihedral angles in Table 2.

Table 1. Since L(3,3)-(6,0) and I(3,3)-(6,0) possess the same number of atoms ($C_{114}H_{12}$), the relative stability can be deduced directly from their total energies, suggesting that L(3,3)-(6,0) is more stable than I(3,3)-(6,0). This method, however, is not suitable for comparison of L(4,4)-(8,0) ($C_{156}H_{16}$) and I(4,4)-(8,0) ($C_{152}H_{16}$), because of the different number of carbon atoms. Here, an alternative method was used to subtract the energy of four carbon atoms far from the defect region, which may be regarded as the atoms in the corresponding pristine tubes. To estimate the average energy of each carbon atom in the pristine (4,4) and (8,0) tubes, the energy of the same tube but with different lengths was computed, respectively. The average energy can then be obtained by the energy difference divided by the difference of the atomic numbers. To compare L(4,4)-(8,0) and I(4,4)-(8,0) with the same number of atoms, the energy of the former with 152 carbon atoms was calculated by subtracting the energy of two carbon atoms of the (4,4) tube and two of the (8,0) tubes from the original energy of L(4,4)-(8,0) with 156 carbon atoms. The computation result is marked with an asterisk in Table 1, showing that the energy of L(4,4)-(8,0) is lower than the corresponding I-IMJ, identical to the comparison result of two (3,3)-(6,0) IMJs.

Geometric Structures. The geometric structures of four IMJs are analyzed to explore the differences between L- and I-IMJs resulting from the different defect distribution. The results show that to accommodate the continuing lattice between two segments, the defect rings and their surrounding hexagons all have some distortions. Especially, remarkable differences are found to exist between heptagons in the different junctions. Therefore, the structures of heptagons are discussed. To coarsely estimate the distortion extent of the heptagons, the bond length and the dihedral angle of the heptagons were calculated and compared with a model molecule 7-methylene-1,3,5-cycloheptatriene shown in Figure 3a, which is a planar molecule. The results are reported in Table 2, wherein the definition of the labels involved is given in Figure 3a,b. From Table 2, it can be seen that

Table 2. Distances (Å) and Dihedral Angles (°) of the Heptagon Defects in the Optimized IMJs[a]

	a	b	c	d	α	β
L(3,3)-(6,0)	1.49	1.40	1.47	1.43	47.5	42.7
I(3,3)-(6,0)	1.44	1.43	1.44	1.42	11.0	32.2
L(4,4)-(8,0)	1.46	1.42	1.47	1.43	30.1	37.3
I(4,4)-(8,0)	1.43	1.44	1.43	1.43	7.0	26.5
7-methylene-1,3,5-cycloheptatriene	1.47	1.36	1.45	1.36	0	0

[a] See Figure 3 for the definitions of the labels a−d, α, and β.

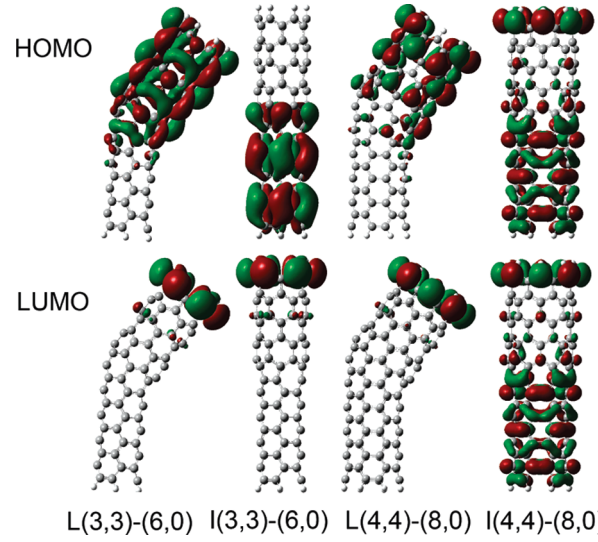


Figure 4. Comparison of the HOMO and LUMO orbitals (an isovalue of 0.01 au) of the L- and I-IMJs at the B3LYP/6-31G(d) level.

compared to the model molecule the difference of the bond length of the heptagon in the L-IMJs is smaller than in the corresponding I-IMJs. The lengths of the seven bonds in the latter tend to be averaged. The distortion of the heptagons in the L-IMJs, however, is found to be more serious than in the I-IMJs by measuring the dihedral angles α and β. Apparently, although the distortion is an unfavorable factor to the stability, the number of defects in the IMJs appears to be more important, resulting in the lower energy of L-IMJs than I-IMJs.

The Characteristic of the Frontier Molecular Orbital. To compare the difference of the frontier molecular orbitals, the spatial distribution of the frontier molecular orbitals and the energy of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) were calculated. The isocontour maps of HOMO and LUMO are depicted in Figure 4. The orbital energy and the HOMO−LUMO gap (ΔE) are listed in Table 1. From Figure 4, an apparent difference of HOMO distribution between L(3,3)-(6,0) and I(3,3)-(6,0) IMJs can be seen. The HOMO of the former is mainly localized on the zigzag section, whereas the HOMO of the latter is localized on the armchair section. The similar phenomenon of localization can also be found in the HOMO orbitals of L(4,4)-(8,0) and I(4,4)-(8,0), but the highest density of HOMO in I(4,4)-(8,0) appears at the zigzag edge. The uneven distribution in one junction can be ascribed to the asymmetry of the structures caused by the defects as well as different chirality, and the difference between L- and

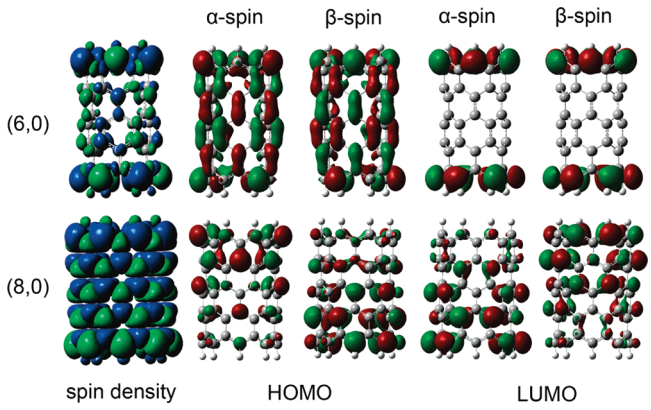


Figure 5. Spin density map of the (6,0) (up left panel) and (8,0) SWCNTs (down left panel) with an isovalue of 0.004 au. Blue color for α-spin electrons, green color for β-spin electrons. α- and β-spin HOMO and LUMO orbitals of (6,0) (up) and (8,0) SWCNTs (down) with an isovalue of 0.02 au.

I-IMJs can be related to the different distribution of the defects. On the other hand, for the distribution of LUMO orbitals, differences only appear to exist between L(4,4)-(8,0) and I(4,4)-(8,0).

From the HOMO−LUMO energy gap ΔE reported in Table 1, the ΔE of L-IMJs is higher than that of the corresponding I-IMJs, indicating that the bent junctions are more stable than the straight ones, identical to the result obtained from the analysis of the total energy. Among the four IMJs, I(4,4)-(8,0) possesses the lowest energy gap, which is also reflected from the similar spatial distribution of its HOMO and LUMO shown in Figure 4.

Spin Polarization. To investigate the possible antiferromagnetic-type spin polarization which will probably change the electronic character of the studied systems, the calculations with the spin-unrestricted approach were carried out. For comparison, the hydrogen-terminated finite-length (6,0) and (8,0) zigzag and (3,3) and (4,4) armchair SWCNTs were also calculated using the same approach. The length of the nanotube, which is defined according to the number of carbon atoms along the tube axis, is selected to be 3.

Among the studied SWCNTs only the (6,0) and (8,0) zigzag nanotubes are found to possess a spin-polarized ground state. Nevertheless, the spin polarization phenomenon has not been observed in I(3,3)-(6,0) and L(3,3)-(6,0) junctions. According to the previous report[21] that spin polarization only appears at the end of the finite-length (n,0) SWCNTs when n is greater than 6; therefore, it is reasonable to conjecture that only very short (6,0) CNTs present the spin-polarized state. To gain further insight into the effect of size on spin polarization, another two (6,0) and (8,0) tubes with the length of 5, same as the length of the zigzag segments in the studied junctions, were also calculated. The results show that the spin polarization only occurred in the zigzag edges of (8,0) SWCNT. As shown in Figure 5, at each side of (6,0) nanotube, the density of α- and β-spin electrons is identical; in sharp contrast, one zigzag edge of the (8,0) segment has a high density of α-spin electrons, while the other edge is rich in β-spin electrons. Moreover, the high-density α- and β-spin LUMO and HOMO orbitals

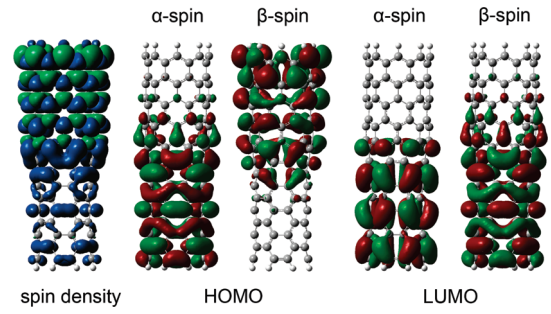


Figure 6. Spin density map of the I(4,4)-(8,0) junction (left panel) with an isovalue of 0.004 au. Blue color for α-spin electrons, green color for β-spin electrons. α- and β-spin HOMO and LUMO orbitals with an isovalue of 0.01 au.

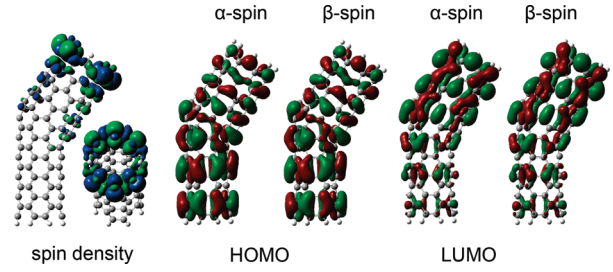


Figure 7. Spin density map of the L(4,4)-(8,0) junction (left panel) with an isovalue of 0.004 au. Blue color for α-spin electrons, green color for β-spin electrons. α- and β-spin HOMO and LUMO orbitals with an isovalue of 0.01 au.

appear on the different sides of the (8,0) nanotube, whereas no difference is observed between the corresponding α- and β-spin orbitals of the (6,0) nanotube. Therefore, the dependence on the diameter and length may result in no spin-polarized ground states in (3,3)-(6,0) junctions, in which the length of the (6,0) segment is 5.

In contrast, both I(4,4)-(8,0) and L(4,4)-(8,0) present a spin polarization ground state. The spin density and the α- and β-spin LUMO and HOMO orbitals are depicted in Figure 6 for I(4,4)-(8,0) and in Figure 7 for L(4,4)-(8,0). In Figure 6, a high density of β-spin electrons is found to appear at the zigzag edge of the junction, while the α-spin electrons are mainly populated around the defect rings at the joint part. Moreover, in the entire armchair segment, the density of α-spin electrons is higher relative to β-spin electrons. It may indicate that the defects induce the spin polarization of the armchair segment. The α-spin HOMO and LUMO orbitals appear to be localized on the armchair segment of the junction. The β-spin HOMO and LUMO orbitals are found to be distributed on the zigzag and armchair segments, respectively. Interestingly, in Figure 7, a local spin polarization in L(4,4)-(8,0) only appears in the radial direction of the zigzag edge. Analysis of the HOMO and LUMO orbitals suggests no marked differences between α- and β-spin states.

Apparently, compared with Figure 4, the effect of spin polarization considerably changed the HOMO and LUMO orbitals of the (4,4)-(8,0) junctions. Whereas no changes occurred in the (3,3)-(6,0) junctions optimized using spin-unrestricted method; therefore, the orbitals are not shown.

In addition, the energies of the four IMJs optimized using the spin-unrestricted method are reported in Table 3. They

Table 3. Results of the Four IMJs at the UB3LYP/6-31G(d) Level[a]

	molecule	energy	spin	HOMO	LUMO	ΔE
L(3,3)-(6,0)	$C_{114}H_{12}$	−4350.555(−0.006)	α	−0.152	−0.120	0.032
			β	−0.152	−0.120	0.032
I(3,3)-(6,0)	$C_{114}H_{12}$	−4350.503(−0.007)	α	−0.152	−0.132	0.020
			β	−0.152	−0.132	0.020
L(4,4)-(8,0)	$C_{156}H_{16}$	−5954.225(−0.021)	α	−0.155	−0.121	0.034
			β	−0.155	−0.121	0.034
I(4,4)-(8,0)	$C_{152}H_{16}$	−5801.712(−0.042)	α	−0.152	−0.109	0.043
			β	−0.161	−0.127	0.034

[a] ΔE is the HOMO-LUMO energy gap. All the energies are in units of Hartree. In Gaussian calculations, the Guess=(Always, Mix) keyword was used, with zero total spin. However, we encountered considerable convergence problems in optimization of I(4,4)-(8,0). Our method to circumvent the problem was to first run the calculation with the 3-21G basis set and then get the initial geometry and guess for the 6-31G(d) basis set calculation from the previous checkpoint file. In the latter calculation, the Guess=(Read, Mix) keyword was used. The energy value in the brackets is the difference between the energy in this table and the corresponding energy in Table 1.

are found to be lower than the corresponding energies using the spin-restricted method at the same level of theory. The preference for I(4,4)-(8,0) is particularly marked. It would be apparent from the table that at variance with the (3,3)-(6,0) IMJs, the spin-polarized HOMO−LUMO gap of the (4,4)-(8,0) IMJs is larger than the corresponding gap in Table 1. These indicate that the spin-polarized states are more stable than the closed shell states for the (4,4)-(8,0) junctions, in which the straight one may be detectable in experiment, and may present half-metallic behavior under the influence of an external electric field.[19−21] Moreover, the α-spin energy gap of I(4,4)-(8,0) is found to be 0.25 eV higher than the β-spin one. This phenomenon is different with the finite-length zigzag nanotubes studied in ref 21, where all the α- and β-spin energy gaps are degenerate in the absence of electric field.

Band Structures of Infinite-Length IMJs. From the above analysis, the existence of hydrogen and defects may induce considerable changes of the electronic properties of the finite-length IMJs studied here. To investigate the effect of the defects on the band structures in the absence of hydrogen atoms, the band structures of the infinite-length IMJs illustrated in Figure 2 have been calculated using periodic boundary conditions. For comparison purposes, the calculations of the corresponding infinite-length pristine SWCNTs have also been carried out. All the results are shown in Figure 8. Comparison of the band structures of the IMJs with the corresponding pristine tubes shows an increase of energy belts near the Fermi energy in the IMJs, which can be related to the defects in the mismatch region. The energies provided by the defect rings or the charge transfer between the different segments lead to the new states near the Fermi energy level. Furthermore, compared to I-IJMs, L-IJMs exhibit a somewhat larger band gap, in accordance with the result of the HOMO−LUMO gap of the finite-length IMJs. Band gap is related to the carrier concentration, which is one of the factors of conductivity. The higher band gaps may imply the possible lower conductivity. Accordingly, our calculations show that the straight junctions may have higher conductivity than the corresponding bent ones. Additionally, from the values of the band gaps shown in Figure 8e−h, the metallicity of the four junctions discussed here appears to have no remarkable dependence on the metallicity of the component segments.

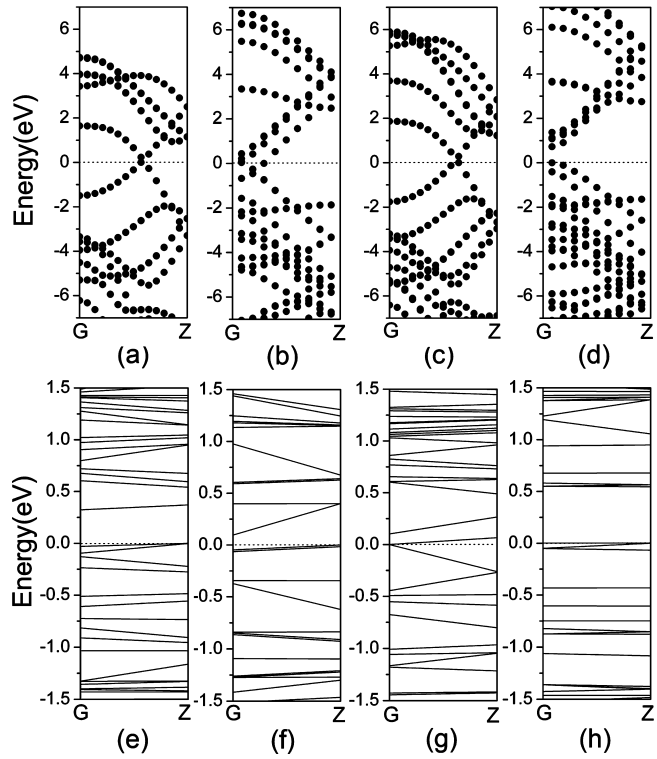


Figure 8. Band structures of the infinite-length pristine SWCNT and the IMJs: (a) (3,3), metallic; (b) (6,0), metallic; (c) (4,4), metallic; (d) (8,0), semiconducting, gap, 0.72 eV; (e) L (3,3)-(6,0), gap, 0.32 eV; (f) I (3,3)-(6,0), gap, 0.10 eV; (g) L (4,4)-(8,0), gap, 0.10 eV; and (h) I (4,4)-(8,0), gap, 0.00 eV. The dashed line denotes the position of the Fermi level.

Conclusions

Bent and straight metallic-metallic and metallic-semiconducting IMJs originating from different defect distribution were constructed and investigated. By analyzing the total energies and the HOMO−LUMO energy gaps of the IMJ structures optimized at the B3LYP/6-31G(d) theoretical level, it has been found that the bent IMJs are more stable than the corresponding straight ones. The spatial distribution of HOMO and LUMO, which is related to the trend of electron transfer, has also been calculated and compared, suggesting a distinct difference between L- and I-IMJs. The spin-polarization phenomenon was observed in the hydrogen-terminated finite-length zigzag SWCNTs but did not occur in the corresponding armchair SWCNTs, which is in agreement with the previous studies. In this contribution, the existence of hydrogen and defects is found to induce the spin polarization in the studied junctions composed of an armchair (4,4) and a zigzag (8,0) nanotube segment, particularly in the straight I(4,4)-(8,0). By analyzing the HOMO and LUMO orbitals, the effect of spin polarization is found to considerably change the electronic properties of the junctions. The spin polarization of the IMJs, however, is shown to be dependent on the diameter and length of the

zigzag segments. Furthermore, the spin-polarized ground state for the I(4,4)-(8,0) IMJ is expected to be detectable.

The effect of the defects, in the absence of the hydrogen atoms, on the band structures of the IMJs is also investigated. The band structures of the infinite-length IMJs have been calculated using periodic boundary conditions. Comparing with the corresponding pristine SWCNTs, more energy states appeared near the Fermi energy in the IMJs due to the asymmetry caused by the defects in the junctions. Moreover, different defect distributions can result in different band structures to further the effect on the conductivity of the IMJs. Our study shows that it may be possible to adjust the electronic properties of an IMJ by constructing the structure with a special defect distribution, which is important in applications.

Acknowledgment. This study is supported by National Natural Science Foundation of China (Nos. 20573102 and 20873066) and also supported by Nankai University ISC.

Supporting Information Available: Geometries of the four studied structures optimized using the B3LYP and UB3LYP with the 6-31G(d) basis set and the four infinite-length IMJs. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chico, L.; Crespi, V. H.; Benedict, L. X.; Louie, S. G.; Cohen, M. L. *Phys. Rev. Lett.* **1996**, *76*, 971–974.

(2) Liu, Y. Q.; Wei, D. C. *Adv. Mater.* **2008**, *20*, 2815–2841.

(3) Treboux, G.; Lapstun, P.; Silverbrook, K. *J. Phys. Chem. B* **1999**, *103*, 1871–1875.

(4) Lee, J. U.; Gipp, P. P.; Heller, C. M. *Appl. Phys. Lett.* **2004**, *85*, 145–147.

(5) Yao, Z.; Postma, H. W. C.; Balents, L.; Dekker, C. *Nature* **1999**, *402*, 273–276.

(6) Li, Y. F.; Hatakeyama, R.; Shishido, J.; Kato, T.; Kaneko, T. *Appl. Phys. Lett.* **2007**, *90*, 173127.

(7) Service, R. F. *Science* **1996**, *271*, 1232–1233.

(8) Li, J. Q.; Zhang, Q.; Chan-Parkb, M. B. *Carbon* **2006**, *44*, 3087–3090.

(9) Charlier, J. C.; Ebbesen, T. W.; Lambin, P. *Phys. Rev. B* **1996**, *53*, 11108–11113.

(10) Ye, Y. F.; Zhang, M. L.; Zhao, J. W.; Liu, H. M.; Wang, N. *J. Mol. Struct.: THEOCHEM* **2008**, *861*, 79–84.

(11) Garau, C.; Frontera, A.; Quinonero, D.; Costa, A.; Ballester, P.; Deya, P. M. *Chem. Phys.* **2004**, *303*, 265–270.

(12) Rochefort, A.; Avouris, P. *Nano Lett.* **2002**, *2*, 253–256.

(13) Son, Y. W.; Lee, S. B.; Lee, C. K.; Ihm, J. *Phys. Rev. B* **2005**, *71*, 205422.

(14) Wu, G.; Li, B. W. *Phys. Rev. B* **2007**, *76*, 085424.

(15) Hua, F.; Fa, W.; Dong, J. *Eur. Phys. J. B* **2005**, *46*, 331–334.

(16) Ouyang, M.; Huang, J. L.; Cheung, C. L.; Lieber, C. M. *Science* **2001**, *292*, 702–705.

(17) Shafranjuk, S. E. *Phys. Rev. B* **2007**, *76*, 085317.

(18) Kim, Y. H.; Choi, J.; Chang, K. J. *Phys. Rev. B* **2003**, *68*, 125420.

(19) Mañanes, A.; Duque, F.; Ayuela, A.; López, M. J.; Alonso, J. A. *Phys. Rev. Lett.* **2008**, *78*, 035432.

(20) Hod, O.; Scuseria, G. E. *ACS Nano* **2008**, *2*, 2243–2249.

(21) Du, A. J.; Chen, Y.; Lu, G. Q.; Smith, S. C. *App. Phys. Lett.* **2008**, *93*, 073101.

(22) 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, Inc.: Wallingford, CT, 2004.

(23) Delley, B. *J. Chem. Phys.* **1990**, *92*, 508–517.

(24) Delley, B. *J. Chem. Phys.* **2000**, *113*, 7756–7764.

(25) Vinciguerra, V.; Buonocore, F.; Panzera, G.; Occhipinti, L. *Nanotechnology* **2003**, *14*, 655–660.

(26) Galano, A.; Orgaz, E. *Phys. Rev. B* **2008**, *77*, 045111.

(27) Bai, L.; Zhou, Z. *Carbon* **2007**, *45*, 2105–2110.

CT900039V

JCTC Journal of Chemical Theory and Computation

Non-Covalent Interactions with Dual-Basis Methods: Pairings for Augmented Basis Sets

Ryan P. Steele,[‡] Robert A. DiStasio, Jr., and Martin Head-Gordon*

Department of Chemistry, University of California, Berkeley, California 94720

Received February 1, 2009

Abstract: Basis set pairings for dual-basis calculations are presented for the aug-cc-pVXZ (X = D, T, Q) series of basis sets. Fidelity with single-basis results is assessed at the second-order Møller−Plesset perturbation theory (MP2) level within the resolution-of-the-identity (RI) approximation, using the S22 set of noncovalent interactions and a series of electron affinities from the G3 set. Root-mean-squared errors for the S22 set are 0.019 kcal mol^{-1} or lower, with a maximum deviation of 0.44%, and errors in nuclear structures are 0.09% or lower. Cost savings of 60−93% (RI-MP2 energies) and 50−88% (RI-MP2 gradients) are demonstrated. Spin-component-scaled MP2 [SCS(MI)-MP2] scaling parameters are provided for the aug-cc-pVXZ series, and dual-basis results are shown to be consistent without reoptimization of the single-basis parameters. Explicit handling of linear dependence in the basis set projection scheme is also provided. These dual-basis pairings will be helpful for accelerating accurate Hartree−Fock, density functional theory (DFT), MP2 and scaled MP2, and so-called doubly hybrid DFT calculations of intermolecular interactions (and other systems), where augmented basis sets are physically important.

1. Introduction

Noncovalent interactions, encompassing $\pi-\pi$ stacking and dispersion forces, hydrogen-bonding, multipole−multipole interactions, and combinations thereof, have proven to be extremely difficult properties to quantitatively predict. Electrostatics alone cannot account for the existence of most of these phenomena, and thus, accurate electron correlation must be included to describe them properly. Further complicating matters, most wavefunction-based methods that properly account for electron correlation are extremely sensitive to the underlying atomic orbital (AO) basis.[1] These basis sets typically require functions with high angular momentum to account for polarization within and between molecules, as well as diffuse exponents to properly account for the long-range nature of the interactions. This combination of accurate electron correlation and large basis sets pushes the capability frontiers of modern electronic structure theory and has limited application to small prototype systems.

Density functional theory (DFT) within the Kohn−Sham formalism[2] has become the de facto method of choice for many theoretical applications. By combining parametrized electron correlation with a cost roughly equivalent to Hartree−Fock (HF), DFT typically performs quite well for thermochemistry[3−7] and molecular structures.[5−8] However, DFT functionals employing the local density approximation (LDA)[9] do not account for the inherently nonlocal effect of dispersion.[10] Even gradient-corrected (GGA)[11−16] functionals are corrections to the local density and do not treat dispersion forces properly. Empirical dispersion corrections (DFT-D)[17−20] have recently gained favor and perform remarkably well for interaction energies, sometimes approaching highly accurate coupled-cluster results. Such terms are purely empirical, however, and have their limitations, such as the neglect of the response of the electron density to the C_6 term. First-principles nonlocal dispersion functionals are also under development and show promise.[21−24]

Another successful method for interaction energies is symmetry-adapted perturbation theory.[25] By constructing the interaction energy between two (or more) distinct subunits directly, expensive computations on the entire system are

* To whom correspondence should be addressed. E-mail: mhg@bastille.cchem.berkeley.edu.

‡ Current address: Department of Chemistry, Yale University, 225 Prospect St., New Haven, CT 06405.

10.1021/ct900058p CCC: $40.75
Published on Web 04/30/2009

unnecessary. Results are typically accurate, and a recent study by Szalewicz[26] demonstrated that, when combined with DFT for the monomers' energies and orbitals, SAPT-DFT[27−32] could effectively and efficiently describe the entire potential energy surface of the benzene dimer, a well-studied prototype system[33−39] known to require extremely accurate electron correlation. These methods also allow for the decomposition of interaction energies into contributions for electrostatics, dispersion, etc., a key tool in determining the factors controlling an interaction. These methods are also not without their shortcomings, however. Applicability is somewhat limited; assessment of conformational energy differences[40] in polypeptides, proteins, or any system with *intra*molecular bonding would not be feasible in the absence of distinct subsytems. Furthermore, a description of chemistry (i.e., broken or changing bonds) is, as yet, undefined in these models.

Thus, an *ab initio* description of electron correlation with the supermolecule approach is still often necessary. The simplest treatment of correlation is second-order Møller−Plesset perturbation theory (MP2).[41] Coupled with the resolution-of-the-identity (RI) approximation,[42−45] MP2 provides an accurate estimate (∼90%) of correlation energies at a cost lower than the underlying self-consistent field (SCF) for many systems of interest. Unfortunately, the same practical shortcoming of all correlated methods also applies to MP2 theory: correlation energies are slowly convergent with respect to the underlying AO basis set.[1] Finally, it should be mentioned that scaling the spin components of the MP2 energy[46−53] has been demonstrated to give improved accuracy, in a statistical sense, both for covalent bond/reaction energies[46,48−51] and intermolecular interactions.[47,52,53] Unfortunately, significantly different scalings are required for these two different classes of interactions. The applicability of these methods in the current context will be discussed in section 3.1.4.

An interesting, practical amalgam of ab initio and DFT calculations has also recently emerged. So-called "doubly hybrid" density functionals[54−60] attempt to correct the poor performance of local functionals for nonlocal properties by including an orbital-dependent, MP2-type term, ideally without overcounting electron correlation or sacrificing accuracy in local properties. Several successful attempts at these qualities have appeared. The most important property of these methods in the present context, however, is that these methods, owing to the wavefunction-like correction, are once again strongly basis set dependent. They are, therefore, also prime candidates to benefit from the basis set pairings presented here.

The dual-basis (DB) method[61−67] has proven to be an accurate alternative to large basis set calculations and provides computational savings of 90% in this regime. In short, the DB scheme involves a standard, iterative SCF calculation in a small subset of the larger, target basis. A perturbative correction (amounting to an approximate Roothaan step) is applied to capture basis set relaxation effects. This scheme defines both DB-HF and DB-DFT. In addition, subsequent correlation corrections could also be added; doing so within the second-order perturbation theory framework defines the DB-MP2 method that is the main focus of this work.

The DB-RI-MP2 method can efficiently produce correlated-electron calculations approaching the AO basis set limit. We have previously presented basis set pairings for 6-31G*,[67] 6-311++G(3df,3pd), cc-pVTZ, and cc-pVQZ.[62] A practical need remains to provide pairings for the aug-cc-pVXZ (X = D,T,Q) series of Dunning-style basis sets,[68−70] commonly used for these noncovalent interactions as well as anionic systems. These basis sets are well-suited to cases where long-range interactions are present and where significant "in−out" flexibility is required in the basis. Frequently, augmented double-ζ results are of comparable quality with the much more expensive nonaugmented quadruple-ζ results for these systems; saturating the diffuse space often leads to faster convergence than higher angular momentum. Furthermore, the aug-cc-pV(D,T,Q)Z series is well-suited to extrapolation schemes,[71,72] with which results may be obtained at the equivalent of one higher angular momentum in the basis set, at essentially no cost. Dual-basis subsets for these bases are constructed and tested herein on several such systems. The main conclusion of this work is that subsets exist that faithfully reproduce target basis quantities, with cost savings equal to or greater than those previously seen for more compact basis sets.

Worth mentioning is a complementary set of correlated ab initio theories. The explicitly correlated methods, typically termed MP2-R12 (or MP2-F12) methods[73] in the context of MP2 theory, exploit the fact that correlation energies converge more slowly than SCF energies with respect to basis set size. These methods are also, in a sense, dual-basis methods; the correlation energies are computed in an additional basis of product functions, which provide basis-set-limit correlation energies. Recent efficient implementations with density fitting,[74,75] local approximations,[76,77] and the RI approximation[78−80] to the costly three- and four-electron integrals are working toward making these methods practical for large systems, and recent examinations of computational cost are very encouraging. Our experience indicates that double- or triple-ζ basis (where R12 energies are often calculated) SCF energies are still not fully converged with respect to basis set, indicated by the fact that the dual-basis correction is nonzero. Therefore, the R12 idea is an exciting correlation counterpart of the dual-basis idea, and future combination of the two methods would be even more efficient than either one individually.

We also note that in Wolinski and Pulay's demonstration of DB-MP2,[63] two truncations of the large aug-cc-pV5Z basis were presented. While this work focuses on the more pragmatic aug-cc-pV(D,T,Q)Z series, it is a testament to DB methods that aug-cc-pV5Z calculations are attainable. After discussing the design of our chosen basis set pairings, this larger basis set's truncation will be discussed in context.

2. Design of Basis Set Pairings

The dual-basis SCF (DB-SCF) method consists of a first-order approximation to basis set relaxation effects and is described in detail in refs 61 and 64 with available extensions

to RI-MP2[62] and its analytic gradient.[65] In short, an iterative self-consistent field (SCF) calculation is performed to convergence in a subset of the larger, target AO basis set, symbolically represented as ⟨target⟩ ← ⟨small⟩, as in 6-311G* ← 6-311G. A single Fock matrix F is constructed in the large basis and is subsequently diagonalized. The resulting molecular orbital (MO) coefficients are then used to form the DB energy correction to the SCF energy, as well as the subsequent correlation calculation:

$$E_{\mathrm{MP2}}^{\mathrm{dual}} = (E_{\mathrm{SCF}}^{\mathrm{small}} + \Delta E_{\mathrm{SCF}}^{\mathrm{target}}) + \Delta E_{\mathrm{MP2}}^{\mathrm{target}} \quad (1)$$

where

$$\Delta E_{\mathrm{SCF}}^{\mathrm{target}} = Tr(\Delta P \cdot F[P]) \quad (2)$$

Here, $\Delta P = P' - P$ is the relaxation of the density matrix upon diagonalization of the single large-basis Fock matrix, $F[P]$.

Savings are most significant when the smaller basis set is a strict subset of the target basis because of integral screening during the large Fock build, as well as during several steps in the analytical gradient. Forming subsets with a sufficiently small basis set ratio (small:target) that still preserve the high accuracy for which these basis sets are intended is the central design challenge.

The aug-cc-pVXZ series is not comprised of sequential supersets; therefore, manual construction of viable subsets is necessary, as was done for cc-pVTZ and cc-pVQZ.[62] Much like the target basis sets themselves, construction of the subsets is inherently empirical but may be guided by judicious chemical insight. Using these same tenets, we have constructed the strongest feasible truncations to the aug-cc-pVXZ series, with the lone restriction that DB results faithfully reproduce full-basis results. Several truncations were considered and tested for each target basis, and results are presented in section 3.

Accurate calculation of properties of noncovalent systems or anionic systems is typically the goal when utilizing these very diffuse basis sets because differential basis set effects (and differential electron correlation, which is strongly basis set dependent) are often crucial. Accordingly, test sets of these systems have been used to guide our empirical construction of subsets. Specifically, we have used the following two test sets for benchmarks:

• The S22 set[81] of Jurecka et al. consists of 22 noncovalent dimer systems, for which either accurate experimental or coupled-cluster binding energies are known. The set consists of 8 dispersion-dominated complexes, 7 hydrogen-bonding complexes, and 7 complexes containing a mixture of these two interactions. This set allows for the accurate parametrization of subsets for both heavy and hydrogen atoms.

• A 25-molecule/atom subset of the G3 set[82−84] is used to calculate adiabatic electron affinities (EA). In both cases, the resultant energy differences are highly sensitive to basis set quality and provide stringent tests of our truncations. The focus of this work is methodology for noncovalent systems. As such, the results for the S22 set have guided our choice of truncation schemes; results for the EA set using these pairings are provided for completeness.

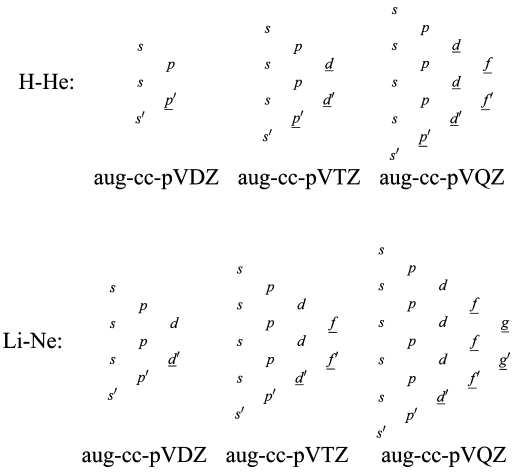


Figure 1. Structure of the aug-cc-pV(D,T,Q)Z series for first- and second-row atoms. The most compact functions are listed at the top of each set, and primed functions depict aug (diffuse) functions. Underlined functions are those eliminated in the truncated dual-aug-cc-pV(D,T,Q)Z series.

The energies of all systems were computed at the RI-MP2 and DB-RI-MP2 levels of theory, using fixed geometries. In all cases, SCF calculations were converged to a DIIS[85,86] error of at most 10^{-8} a.u., using integral thresholds of 10^{-12} a.u. For the RI calculations, the corresponding auxiliary basis sets of Weigend[87] were used. Linear dependence was handled as described in Appendix A, using a drop tolerance of 10^{-6}. The frozen core approximation was used in all correlated (RI-MP2) calculations.

The balanced structure of the aug-cc-pV*X*Z basis sets is shown in Figure 1. Lessons learned from previous DB truncations[62,67] help guide the choice of subsets. In particular, some polarization functions are required in the small basis for quantitative agreement with target basis properties.[67] The new design challenge for the augmented basis sets is, thus, choosing which diffuse functions may be discarded, as well as the proper balance between polarization and diffuse functions. All viable candidates for truncations of aug-cc-pV(D,T)Z are given in the Supporting Information (SI) and categorized according to the truncation level on heavy and hydrogen atoms.

The qualitative standards we sought to satisfy are as follows:

(1) The error caused by using the DB method should be sufficiently less than the difference between basis sets (i.e., between aug-cc-pVDZ and aug-cc-pVTZ).
(2) The accuracy should be balanced across anions, dispersion complexes, and hydrogen-bonded complexes.
(3) The cost savings resulting from the truncation should be on par with, or greater than, previous DB pairings (∼90−95% for RI-MP2 energies).

Because of the cost and number of possible permutations, the results for aug-cc-pVDZ and aug-cc-pVTZ were used to guide seven possible truncations for the aug-cc-pVQZ basis, also provided in the Supporting Information.

Table 1. DB-RI-MP2 Basis Pairing Results for aug-cc-pV(D,T,Q)Z on the S22 Set[81] [a]

	basis	relative to	CH ratio[b]	non-CP[c] rms[d]/ kcal mol^{-1}	CP rms/ kcal mol^{-1}
dual	aug-cc-pVDZ ← dual-aug-cc-pVDZ	aug-cc-pVDZ	0.750	0.085	0.043
	aug-cc-pVTZ ← dual-aug-cc-pVTZ	aug-cc-pVTZ	0.536	0.034	0.019
	aug-cc-pVQZ ← dual-aug-cc-pVQZ	aug-cc-pVQZ	0.357	0.020	0.015
single	cc-pVDZ	aug-cc-pVDZ		2.629	2.151
	cc-pVTZ	aug-cc-pVTZ		0.740	0.821
	aug-cc-pVDZ	aug-cc-pVTZ		1.462	0.730
	aug-cc-pVTZ	aug-cc-pVQZ		0.756	0.269

	basis	relative to	C_{OS}/C_{SS}[e]	non-CP rms/ kcal mol^{-1}	CP rms/ kcal mol^{-1}
dual	aug-cc-pVDZ ← dual-aug-cc-pVDZ	CCSD(T)/CBS		3.76	1.20
	aug-cc-pVTZ ← dual-aug-cc-pVTZ			2.29	1.04
	aug-cc-pVQZ ← dual-aug-cc-pVQZ			1.57	1.10
	aug-cc-pVDZ ← dual-aug-cc-pVDZ		0.00/1.83		0.26
	aug-cc-pVTZ ← dual-aug-cc-pVTZ		0.22/1.52		0.26
	aug-cc-pVQZ ← dual-aug-cc-pVQZ		0.31/1.40		0.26
single	cc-pVDZ	CCSD(T)/CBS		1.88	2.49
	cc-pVTZ			1.72	0.91
	cc-pVQZ			1.46	0.90
	aug-cc-pVDZ			3.73	1.19
	aug-cc-pVTZ			2.29	1.06
	aug-cc-pVQZ			1.57	1.11
	aug-cc-pVDZ		0.00/1.83		0.26
	aug-cc-pVTZ		0.22/1.52		0.26
	aug-cc-pVQZ		0.31/1.40		0.26

[a] The results shown in the upper section are errors (kcal mol^{-1}) relative to calculations in the basis shown in the second column, and those shown in the lower section are errors (kcal mol^{-1}) relative to complete basis set (CBS) estimates of coupled-cluster singles and doubles with perturbative triples (CCSD(T)). [b] Basis set size ratio for a molecule containing an equal number of heavy and hydrogen atoms. [c] rms = Root mean squared deviation. [d] CP = Counterpoise-corrected. [e] SCS(MI) scaling coefficients for same-spin (SS) and opposite-spin (OS) components of the MP2 correlation energy, refit to aug-cc-pV(D,T,Q)Z data using the method of ref 52.

Table 2. DB-RI-MP2 Basis Pairing Results for aug-cc-pV(D,T,Q)Z on a 25-Molecule Electron Affinities Subset of the G3 Set[82−84a]

	basis	relative to	rms[b]/kcal mol^{-1}
dual	aug-cc-pVDZ ← dual-aug-cc-pVDZ	aug-cc-pVDZ	0.109
	aug-cc-pVTZ ← dual-aug-cc-pVTZ	aug-cc-pVTZ	0.143
	aug-cc-pVQZ ← dual-aug-cc-pVQZ	aug-cc-pVQZ	0.261
	aug-cc-pVDZ ← dual-aug-cc-pVDZ	expt	6.256
	aug-cc-pVTZ ← dual-aug-cc-pVTZ	expt	5.397
	aug-cc-pVQZ ← dual-aug-cc-pVQZ	expt	5.549
single	cc-pVDZ	aug-cc-pVDZ	29.65
	aug-cc-pVDZ	aug-cc-pVQZ	3.907
	aug-cc-pVTZ	aug-cc-pVQZ	1.162
	aug-cc-pVDZ	expt	6.238
	aug-cc-pVTZ	expt	5.343
	aug-cc-pVQZ	expt	5.501

[a] Results are errors (kcal mol^{-1}) relative to the method shown in the second column. [b] rms = Root mean squared deviation.

3. Results

Optimal basis set pairing schemes are depicted in Figure 1. Summarized results for the S22 and EA sets are shown in Tables 1 and 2 for these pairings, along with pertinent basis set size information. A full statistical analysis of these results and the raw data are included in the Supporting Information for all tested pairings. The results for each basis set (D,T,Q) are here briefly discussed in turn.

3.1. Energies. *3.1.1. aug-cc-pVDZ.* The aug-cc-pVDZ basis[68] has a relatively limited number of feasible truncation schemes, but unlike 6-31G*,[88−90] three sets of diffuse functions are available for truncation. As the aug-cc-pVXZ series is often considered overly diffuse,[91] these diffuse functions are viable candidates for elimination. The removal of the diffuse d' functions (and p' functions on hydrogen) serves as a useful first attempt. Results for this pairing are quite good, as root-mean-squared (rms) errors are only 0.085 kcal mol^{-1} for the S22 set and 0.109 kcal mol^{-1} for electron affinities, relative to aug-cc-pVDZ results. The same errors for a single-basis cc-pVDZ calculation are 2.629 and 29.654 kcal mol^{-1}. Alternative elimination of the polarization functions, while retaining the diffuse functions, leads to significantly worse DB results (see Supporting Information). Given that the cost of this second pairing is actually greater

than the cost of the first, because of the inevitably larger number of significant shell pairs in the small basis, this pairing was eliminated. Once again, we find that retention of at least some polarization is necessary for accurate replication of target-basis properties. For completeness, a more drastic truncation was constructed, in which the two sets of diffuse functions of highest angular momentum were eliminated. This pairing performs the worst of the three, despite retention of polarization functions in the small basis. The lone set of diffuse s' functions on heavy atoms is not sufficient to account for these diffuse properties; in this small basis, even the aug functions serve the role of polarization functions. These trends appear to be constant for heavy and hydrogen atoms. No mixing of these pairing combinations performs better. Thus, we suggest the first pairing for use with aug-cc-pVDZ calculations, shown in Figure 1.

3.1.2. aug-cc-pVTZ. The scope for computational savings without loss of accuracy is greater at the aug-cc-pVTZ level and will be the main focus of our discussion. Using the previous 4s3p2d truncation of cc-pVTZ as a guide, the most logical extension was the additional elimination of the diffuse f' functions. Errors are small for this pairing, only 0.015 kcal mol^{-1} error for the S22 set. Three further truncations were attempted by removing one set of d functions, all of which lead to greater savings. Results for all of these pairings demonstrate rms errors below 0.08 kcal mol^{-1} for the S22 interaction energies. While removal of the central d performs nearly as well as the more costly f-only truncation, a more cost-effective option is available. Removal of the diffuse d' function performs just as well on the S22 set and nearly four times better for electron affinities. Thus, while augmentation of the basis set is crucial for electron affinities, accurately capturing polarization effects is also a necessary requirement. This choice is also entirely consistent with the previous nonaugmented cc-pVTZ truncation; for the augmented set presented here, an additional diffuse function has simply been removed.

A similar trend is observed for hydrogen, even in hydrogen-bonded complexes. Thus, the 4s2p truncation is the truncation of choice for aug-cc-pVTZ for hydrogen, in which the diffuse p' function was eliminated. In fact, in all of the optimal sets chosen here, diffuse functions beyond s' were found to be unnecessary for hydrogen.

For this pairing, errors are sufficiently small that dual-basis calculations may serve as a viable replacement for full-basis calculations of intermolecular interactions. CP-corrected rms errors in the S22 set are 0.019 kcal mol^{-1} (max 0.042 kcal mol^{-1}), an average absolute percent error of only 0.25% (max 1.17%). They are also significantly below the errors resulting from the use of aug-cc-pVDZ, one of the main requirements of our truncations. For example, the same errors from using single-basis aug-cc-pVDZ calculations are 0.730 kcal mol^{-1} (max 1.563 kcal mol^{-1}) and 10.3% (max 19.6%).

The results for adiabatic EAs are also reasonable. The rms error (relative to aug-cc-pVTZ) is 0.143 kcal mol^{-1}. The errors relative to experimental EAs for dual- and single-basis calculations are 5.40 and 5.34 kcal mol^{-1}, respectively. Even aug-cc-pVQZ remains 5.50 kcal mol^{-1} from experimental values, indicating that the greater influence in these errors is the MP2 model itself, rather than basis set effects. In this context and given the performance for S22, this pairing is quite adequate.

3.1.3. aug-cc-pVQZ. As is the standard story with dual-basis calculations, both savings and accuracy increase as the target basis becomes more complete. The same is seen for aug-cc-pVQZ here. Three main pairings were considered: a conservative truncation of the two highest angular momentum levels ($2f/f'$ and g/g'), a subsequent removal of diffuse functions of the next lower angular momentum (d'), and a final pairing in which one polarization and one diffuse function were retained in the d level of heavy atoms and the p level of hydrogen atoms. In the latter pairing, the "middle" polarization function was retained. Other permutations of this latter truncation could be performed, but given the expense of these calculations, this balanced truncation seems appropriate. Preliminary tests of more drastic truncations (not shown) proved to be significantly worse.

The least aggressive pairing expectedly performs best, with S22 rms errors of 0.013 kcal mol^{-1} (max 0.037). The estimated SCF savings (see section 3.3) of 90% is already significant, but further truncation is possible. Removal of the next set of diffuse functions produces errors of only 0.015 kcal mol^{-1}, with 93% savings. Thus, the sets of s' and p' diffuse functions are capable of capturing diffuse properties, a fact consistent with the diffuse function scheme in Pople-style basis sets.[92] Alternatively, the last tested pairing includes one set of d polarization functions (the middle of the three) and one set of diffuse d' functions. This set accounts for the aforementioned in−out flexibility and retains some diffuse character, with a CP-corrected rms error of only 0.016 kcal mol^{-1}. However, as was discussed for aug-cc-pVTZ above, a more cost-effective option is the second choice, in which the three polarization functions are retained and the d' diffuse function is eliminated. While this choice leaves three, as opposed to two, d functions in the smaller basis set, the diffuse d' function contributes more significantly to the overall cost (a fact not accounted for in the simple basis set scaling ratios). This choice is also more consistent with the previously published cc-pVQZ truncation. For hydrogen atoms, only s and p functions were required, and only the diffuse s' must be retained.

3.1.4. Spin-Component-Scaled MP2. To correct for deficiencies in MP2 theory, such as overestimating dispersion and poor treatment of atomization energies, Grimme[46] introduced two empirical parameters to the MP2 model, which separately scale the opposite-spin (singlet, OS) and same-spin (triplet, SS) components of the correlation energy. These methods have been termed spin component scaled (SCS) methods. While several flavors of SCS have since been introduced,[48−50] DiStasio[52] recently noted that the optimal scaling of OS and SS components for noncovalent systems is opposite the optimal relative scaling for thermochemical properties. A basis set-dependent set of scaling parameters was suggested for molecular interactions (MI) and termed SCS(MI); the resulting methodology was recently shown to perform the best among scaled MP2 methods for the uracil dimer.[93] Thus, a globally optimized set of empirical parameters appears not to exist, most likely because the empirical

scaling is accounting for different decay properties of the components of the correlation energy. For a well-defined class of systems, however, such as the S22 set considered here, this scaling is reasonable.

In Table 1, SCS(MI) results for single- and dual-basis interaction energies are shown. The original set of scaling parameters was optimized for the (nonaugmented) cc-pV(D,T,Q)Z series of basis sets. We have thus performed the same fitting of coefficients for the single-basis aug-cc-pV(D,T,Q)Z series here; scaling coefficients for OS and SS are also shown in Table 1. The optimal scaling parameters are markedly basis set-dependent (ranging from complete omission of OS components for aug-cc-pVDZ to $c_{OS} = 0.31$ for aug-cc-pVQZ), in contrast to Grimme's original assertion that SCS parameters are independent of basis set for thermochemical properties. The errors, relative to complete-basis-set (CBS) estimates of CCSD(T) binding energies, however, are remarkably basis set-independent when scaled appropriately. In fact, unscaled interaction energy errors also appear to be roughly independent of basis set when CP-corrected, as long as diffuse functions are included in the basis. Nonaugmented cc-pVDZ, for example, performs notably worse.[47,52]

Overall, SCS(MI) results are good for all basis sets considered. The across-the-board rms error of 0.26 kcal mol^{-1}, relative to CCSD(T), is a worthy improvement over the 1.1−1.2 kcal mol^{-1} error seen for unscaled energies. Importantly, dual-basis results are consistent with single-basis results, using the same scaling parameters optimized for single-basis energies.

3.2. Structures. Design of the pairings established above utilized fixed-geometry binding energies as the metric of interest. Structural optimization, however, is an important component for these noncovalent complexes; the recently developed DB-RI-MP2 analytical gradient[65] provides an efficient means for this additional comparison. For practical reasons, an 11-molecule subset of the S22 set was tested, which contains the systems with, at most, one aromatic ring. Again using single-basis MP2/aug-cc-pVQZ results as the benchmark of interest, the (mass-unweighted) center-to-center distances of the 11 molecules were tabulated with the aug-cc-pVXZ series and their DB counterparts. Such a metric is not unique but most directly describes the intermonomer aspect of the structural optimization. Nuclear geometries were optimized to a maximum gradient component of 3×10^{-5} a.u. and either a displacement of 12×10^{-5} a.u. or an energy change of 1×10^{-6} a.u., an order of magnitude tighter than the Q-Chem[94] default tolerances, since the potential energy surfaces involved are relatively flat.

Summarized results are presented as rms errors in Figure 2. The first noteworthy trend is the sizable basis set dependence of the structures. While intramolecular geometries are often less basis set dependent, the intermonomer spacing in these noncovalent complexes relies almost wholly on differential electron correlation and is a stringent test of the methodology presented here. The single-basis cc-pVDZ results, for example, are in error by more than 0.1 Å. Augmentation of the basis set reduces this error by better than a factor of 2. In fact, aug-cc-pVDZ results are, on average, more than a third better than (nonaugmented) cc-pVTZ structures. Even cc-pVQZ produces structures that are still 0.026 Å from structures of its augmented counterpart, and again, the augmented basis set of one lower angular momentum (aug-cc-pVTZ) outperforms the nonaugmented basis.

The DB-RI-MP2 results, obtained significantly faster, as discussed in the following section, are generally consistent with the single-basis results that they were designed to mimic. The rms errors for single- and dual-basis aug-cc-pVDZ are 0.041 and 0.042 Å, respectively. While the overall errors do not exactly reproduce aug-cc-pVQZ results, this augmented double-ζ pairing is the most likely basis set to be utilized for structural optimizations. The excellent fidelity of the DB results with target-basis quantities indicates that this pairing is a viable means for geometry optimizations. Results for DB aug-cc-pVTZ and aug-cc-pVQZ each show further significant improvements in accuracy, although the DB errors relative to the target basis are larger than those for aug-cc-pVDZ. While this observation is distinct from previous results for covalently bound systems,[62,64] it may also be expected from the progressively more aggressive truncation schemes presented herein. Note that the DB error is roughly consistent, however, with the previously published cc-pVQZ pairing. Furthermore, while the DB errors seen here are larger than those demonstrated for covalent systems, the intermonomer spacings (ranging from 1.6 to 3.7 Å) are also significantly larger than typical intramolecular bonds. The largest DB error (aug-cc-pVQZ), for example, is still within an average unsigned error of 0.085% of target-basis separations and is significantly less than the intrinsic error resulting from using aug-cc-pVTZ (0.464%).

The interaction energies of these optimized complexes are presented in the lower panel of Figure 2. Again, strong basis set dependence is demonstrated. The effect of basis set augmentation is less pronounced than in the structures but is still significant. Consistency between single- and dual-basis results is demonstrated, with differential errors in the augmented pairings of 0.03 kcal mol^{-1} or less.

3.3. Timings. Representative nuclear force timings are presented in Figure 3. All timings were performed on a single 2 GHz Opteron processor, using the same calculation parameters described above. The timings in the figure are broken down into constituent contributions.

For a molecule with an equal number of heavy and hydrogen atoms, the basis set truncation ratio for the aug-cc-pVDZ pairing is 0.750, leading to an estimated cost savings of 60% in the SCF calculation (see ref 62 for cost estimation formulas; 12 SCF cycles have been assumed for estimation purposes). For single-point energies, this savings is significant. For analytic SCF (Hartree−Fock or density functional theory) gradients, this truncation simply is not drastic enough to produce more than modest savings (28%). Using the recently developed DB-RI-MP2 analytical gradient,[65] however, we expect to see somewhat more significant savings. While the savings are system size-dependent, a simple timings comparison for the

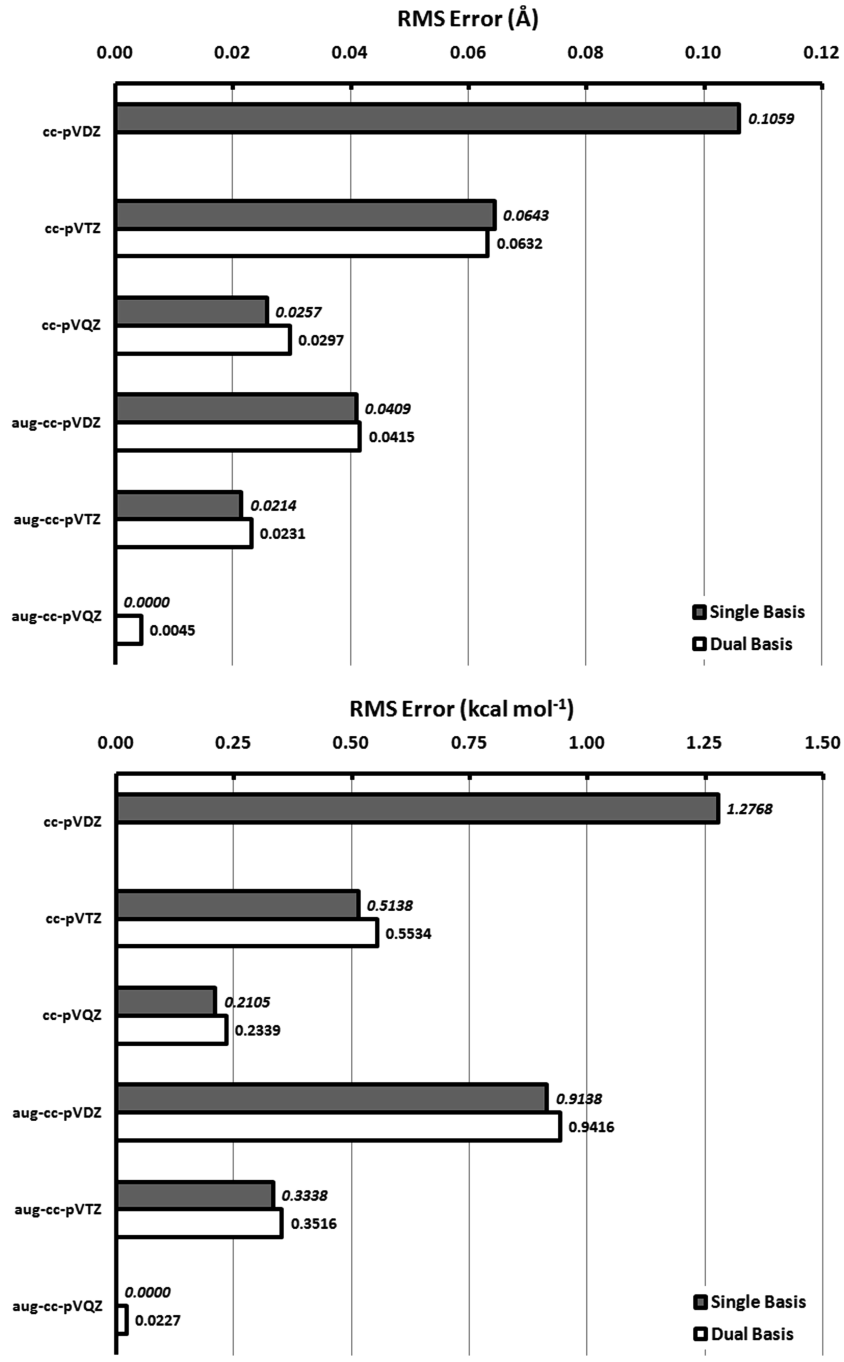


Figure 2. Statistical summary of DB-RI-MP2 performance on an 11-molecule subset of the S22 set. Shown are rms errors in optimized center-to-center distances (top), as well as errors in non-counterpoise-corrected interaction energies (bottom) on these optimized structures. Both panels utilize MP2/aug-cc-pVQZ results as reference numbers and present single- and dual-basis errors.

benzene dimer shows that total nuclear force savings are indeed 50%. Interaction energy errors using aug-cc-pVDZ are typically significant (though much improved over nondiffuse basis sets, such as cc-pVDZ or 6-31G*), relative to the basis set limit, but structures are often accurate and may be the only currently viable option for biologically relevant molecules.

Unlike the truncation for aug-cc-pVDZ, the estimated timings for aug-cc-pVTZ SCF nuclear gradients show significant promise. At the DB-SCF level, the basis truncation ratio of 0.536 leads to a 65% savings in a total nuclear force calculation. Thus, dual-basis aug-cc-pVTZ geometry optimizations may be performed three times faster than their single-basis counterparts. Further truncations may be feasible for geometry optimizations, but here we choose to use energies as the benchmark of interest. Additionally, savings in the DB-SCF gradient are somewhat tempered by the need to solve a response (z-vector)[95] equation. For the DB-RI-MP2 gradient, the response equation is already present in single-basis calculations, and the dual-basis savings should be more significant. With benzene dimer, for example, DB-RI-MP2 nuclear

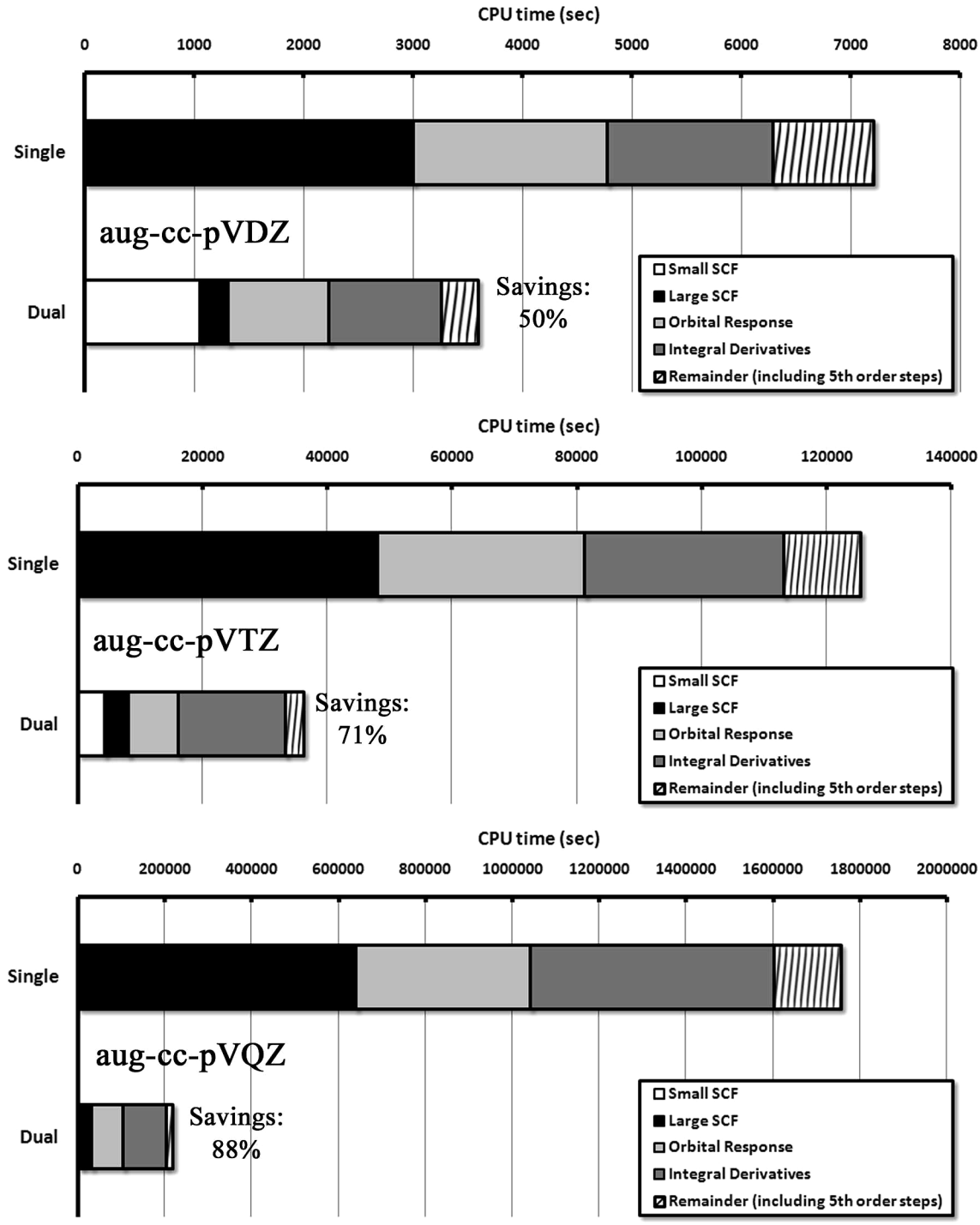


Figure 3. Nuclear force calculation timings for the benzene dimer. Shown are constituent contributions to the nuclear force for single- and dual-basis aug-cc-pV(D,T,Q)Z. Note the vastly different scales.

force timings show a 71% savings. In fact, the total DB-RI-MP2 nuclear force calculation was 25% faster than the SCF alone in the comparative single-basis calculation!

For aug-cc-pVQZ, an estimated 93% reduction in SCF cost is anticipated. Further extension to DB-RI-MP2 nuclear gradients, still admittedly prohibitive for many systems, exhibit noteworthy savings. For the benzene dimer, an 88% reduction in nuclear force timings is observed, reducing a *single* nuclear force calculation from nearly two weeks to two days.

4. Discussion

The results of this brief analysis indicate that DB pairings for augmented Dunning-style basis sets are quite viable. Accuracy closely approaches the results of target basis quantities, at savings of 60−93%. Further analysis indicates that basis set superposition error,[96] an unfortunate artifact of finite-basis calculations, particularly in noncovalent interactions, is slightly larger in DB calculations, but counterpoise-corrected results are just as accurate (see Supporting Information). Performance on noncovalent com-

plexes is superior to performance for electron affinities (in terms of absolute error), and this latter area does expose one limitation of DB methods, within this choice of basis set pairings. However, since anionic and open-shell complexes are, admittedly, a difficult class of systems for MP2 anyway,[97] these errors are tolerable.

Given the above results, the aug-cc-pV5Z truncation of Wolinski and Pulay[63] appears reasonable. The original 7s6p5d4f3g2h basis is truncated to 7s6p4d3f in their best pairing. Based upon our findings for aug-cc-pVQZ, this pairing most likely could be truncated further, either by elimination of all but the central *f* function or by elimination of the *f* functions altogether. Without empirical testing on our part, however, we recommend their original truncation for this extremely large basis set regime.

The SCS(MI) results are encouraging, in that the spin-component scaling is transferrable to DB pairings. A critical assessment of the basis set dependence in the S22 results, however, must include the observations that (1) The CP-corrected *aug* single-basis errors are essentially basis set independent, and (2) the dual- and single-basis results are independent of basis set after spin-component scaling. Both raise the question of why to use these larger basis sets at all. Several justifications exist. First, non-CP results are significantly more basis set dependent, as was shown in Figure 2. The CP correction is only viable for well-defined monomeric units. Intramolecular interactions in biological systems would necessarily suffer the much larger basis set dependence seen in the non-CP column of Table 1. Second, the better performance of medium-sized basis sets is solely caused by error cancellation, where the overestimate of dispersion interactions inherent to MP2 is canceled by basis set incompleteness effects. This error cancellation will not necessarily hold across the entire potential energy surface, where dispersion contributions are variable. And finally, noncovalent systems are often coupled with chemical systems, for example, enzymatic reactions, where larger basis sets more strongly influence thermochemistry. A balanced description of both properties is essential in this class of systems, and large basis sets are still required.

5. Conclusions

An accurate description of noncovalent interactions requires the use of correlated-electron methodologies and large basis sets. This pair of requirements, however, comes with a steep computational cost. The above work has shown that dual-basis MP2 is a viable alternative to single-basis calculations for noncovalent interactions. Optimal basis set pairings for the aug-cc-pV(D,T,Q)Z series were constructed, and, despite relatively aggressive basis truncations, errors for the S22 set are on the order of hundredths of a kcal mol^{-1}, a necessary requirement for a class of systems displaying small energy differences. Because of the ability to truncate the overly complete diffuse space of the augmented Dunning basis sets, computational time savings are significant for all three basis sets considered, and savings grow with the size of the basis. Since correlated-electron calculations converge slowly with respect to basis set size, these results allow for the accurate calculation of nearly complete basis set properties at significantly reduced cost.

While our assessment of both accuracy and computational savings are performed for dual-basis RI-MP2 calculations, we expect that these augmented dual basis sets will be useful in a variety of other electronic structure methods, as well. For instance, if calculations of intermolecular interactions are performed using a density-functional method[17−24,54,55,98,99] or post-Hartree−Fock method[100−103] that includes dispersion corrections, then large augmented basis sets are also important. These DB methods will provide speedups that will be greater than those reported here. Likewise, our conclusions regarding computational savings in MP2 calculations directly transfer to other methods which make second-order perturbative corrections to DFT energies, such as the increasingly popular "double-hybrid" functionals of Grimme.[54,54−59] Calculations using double-hybrid methods converge similarly with basis set size to MP2 itself. Finally, a natural synergism exists between DB methods, which correct SCF energies, and the emerging R12/F12 methods,[73] which provide a dual, two-particle basis for describing electron correlation effects.

Acknowledgment. Funding for this work has been provided by the National Science Foundation under Grant No. CHE-0535710, with additional support for code development from a subcontract from Q-Chem, Inc. for an NIH Small Business Innovation Research (SBIR) grant. M.H.-G. is a part owner of Q-Chem, Inc.

Supporting Information Available: Tested basis set pairings, detailed S22 results, BSSE analysis, detailed electron affinities data, basis set files in Q-Chem format. This information is available free of charge via the Internet at http://pubs.acs.org/.

Appendix A: Linear Dependence

In a dual-basis calculation, a projection of the occupied molecular orbital coefficients is required. The projection from one atomic orbital basis set to its corresponding superset is straightforward, but the presence of linear dependence in one or both basis sets requires attention. This Appendix describes these linear dependence possibilities and their associated solutions. The following methods have been implemented in Q-Chem.[94]

For reference, the following notational conventions are used:

$p, q, r, \ldots$ = all molecular orbitals (MOs)
$i, j, k, \ldots$ = occupied MOs
$a, b, c, \ldots$ = virtual MOs
$u, v, \lambda, \ldots$ = atomic orbitals (AOs)

Unadorned indices represent small-basis quantities. Indices bearing a tilde ($\tilde{\mu}$) represent large-basis-only quantities ("new functions"), while barred indices ($\bar{\mu}$) represent the full large-basis space; that is, $\{\mu\}\oplus\{\tilde{\mu}\} = \{\bar{\mu}\}$. The nonorthogonal AO space may be transformed to an orthogonalized AO (OAO) space by

$$|\underline{\mu}\rangle = \sum_{\nu} |\nu\rangle X_{\nu\underline{\mu}}$$

where underlined indices signify OAO functions. The nonunitary transformation matrix X may be chosen as any matrix that leaves the basis orthogonal; however, two forms are most commonly used.[104] Symmetric orthogonalization employs

$$X = S^{-1/2}$$

where the AO overlap matrix is defined as $S_{\mu\nu} = \langle\mu|\nu\rangle$. Canonical orthogonalization utilizes

$$X = Us^{-1/2}$$

where the unitary U matrix diagonalizes S, giving $USU^T = s$.

If the chosen basis set is linearly dependent, symmetric orthogonalization is not possible, since the S matrix becomes singular and its inverse ill-defined. Canonical orthogonalization is typically employed in this case, where the columns of U are scaled by $s^{-1/2}$. If any eigenvalue is below a chosen numerical threshold, the column is instead eliminated, thus reducing the number of OAOs by the number of linear dependencies and removing the null space.

An MO $|p\rangle$ may, therefore, be expanded in the small-basis AO space as

$$|p\rangle = \sum_{\mu\nu}^{\mathrm{AO}} \sum_{\underline{\lambda}}^{\mathrm{OAO}} |\mu\rangle X_{\mu\underline{\lambda}} X_{\underline{\lambda}\nu} \langle\nu|p\rangle \tag{3}$$

$$= \sum_{\mu} C_{\mu p} |\mu\rangle \tag{4}$$

Starting with this same MO, the large-basis analogue may be constructed via projection

$$|\bar{p}\rangle = \sum_{\bar{\lambda}\bar{\sigma}} \sum_{\mu\nu} \sum_{\underline{\gamma}} \sum_{\underline{\delta}} |\bar{\lambda}\rangle X'_{\bar{\lambda}\,\underline{\delta}} X'_{\underline{\delta}\,\bar{\sigma}} \langle\bar{\sigma}|\mu\rangle X_{\mu\underline{\gamma}} X_{\underline{\gamma}\nu} \langle\nu|p\rangle \tag{5}$$

$$= \sum_{\bar{\lambda}\bar{\sigma}} \sum_{\mu} \sum_{\underline{\delta}} |\bar{\lambda}\rangle X_{\bar{\lambda}\,\underline{\delta}} X_{\underline{\delta}\bar{\sigma}} S_{\bar{\sigma}\mu} C_{\mu p} \tag{6}$$

where $S_{\bar{\sigma}\mu}$ is the (rectangular) overlap matrix in the mixed large-small space. (The primed X's are simply a reminder that different orthogonalization schemes could potentially be used for the two basis sets.) An occupied MO coefficient in the large space is thus constructed as

$$C_{\bar{\lambda}i} = \sum_{\bar{\sigma}} \sum_{\mu} \sum_{\underline{\delta}} X_{\bar{\lambda}\,\underline{\delta}} X_{\underline{\delta}\,\bar{\sigma}} S_{\bar{\sigma}\mu} C_{\mu i} \tag{7}$$

In the absence of linear dependence, several simplifications occur. Equation 3 becomes

$$|p\rangle = \sum_{\mu\nu}^{AO} |\mu\rangle S^{-1}_{\mu\nu} \langle\nu|p\rangle \tag{8}$$

and a projected large-basis occupied MO coefficient reduces to

$$C_{\bar{\lambda}i} = \sum_{\bar{\sigma}} \sum_{\mu} S^{-1}_{\bar{\lambda}\bar{\sigma}} S_{\bar{\sigma}\mu} C_{\mu i} \tag{9}$$

This latter scheme properly handles a projection between any two linearly independent basis sets (activated by the BASISPROJTYPE=OVPROJECTION keywords in Q-Chem). Importantly, when $\{\mu\} \subset \{\bar{\mu}\}$, the only case considered for our basis set truncations (for computational efficiency reasons), the rectangular projection matrix

$$T_{\bar{\lambda}\mu} = \sum_{\bar{\sigma}} S^{-1}_{\bar{\lambda}\bar{\sigma}} S_{\bar{\sigma}\mu} \tag{10}$$

simply becomes a rectangular delta function matrix

$$T_{\bar{\lambda}\mu} = \delta_{\bar{\lambda}\mu} \tag{11}$$

The large-space occupied coefficients are, therefore, identical to the small-space occupied coefficients, with additional null elements corresponding to new basis functions. The same is true for the large-basis density matrix, defined as

$$P_{\bar{\mu}\bar{\nu}} = \sum_{i} C_{\bar{\mu}i} C_{i\bar{\nu}}$$

The small-basis elements of P are identical to elements in its small-basis analogue, p, with all other elements equal to zero.

When two or more basis functions cause the basis to become (numerically) linearly dependent, however, this scheme requires revision. Standard routines to eliminate linear dependence, such as canonical orthogonalization (or equivalently, a "square" singular value decomposition),[105] take linear combinations of the offending basis functions to produce an orthonormal basis of reduced dimension. Several issues make this choice tenuous for a dual-basis projection.

First, while canonical orthogonalization in eq 7 is qualitatively and numerically correct, it, by definition, mixes the contribution of each AO to the large-basis MO coefficients. Thus, the zero-structure in the resultant coefficients vanishes, leaving the integral screening in the large basis ineffective. The addition of an exactly linearly dependent function in the large basis, for example, would produce an occupied MO coefficient with two identical elements, both equal to half of the corresponding element in the small basis.

The second (related) issue involves the large-basis density matrix. When the aforementioned mixing occurs, the small-basis elements of P also change. Thus, the reference energy before projection, $E[p]$, is no longer equal to the energy after projection, $E[P]$, and ambiguity arises in the definition of the dual-basis energy.[62]

This X-dependent ambiguity also arises in a standard SCF calculation and can even leave the SCF energy effectively nonvariational. Consider, for example, a calculation of the energy of a hydrogen atom, for which a single s function's exponent has been variationally optimized. If a second s function causes linear dependence, the resultant "mixed" s function will necessarily cause the energy to rise. While this latter property is rare, the fact remains that the SCF energy is dependent on X for linearly dependent basis sets. The same is true for a dual-basis calculation, with the added possibility that the *same* calculation may employ different X.

We thus desire a projection matrix T that accomplishes the following:

(1) The resultant large-basis occupied MO coefficients are orthonormal.
(2) The resultant density matrix remains idempotent.
(3) The corresponding large-basis density matrix retains its designed "zero structure".

(4) The small-basis energy is unchanged by the projection ($E[p] = E[P]$) and is unambiguously defined.

The ansatz for T that we have chosen is simply the delta function projector shown in eq 11, applied whether or not linear dependencies exist. The large-basis occupied MO coefficients are a strict superset of the small-basis coefficients and are used to construct the large basis density matrix P. Canonical orthogonalization is subsequently used for later stages of the dual-basis SCF, such as the transformation to an OAO basis prior to diagonalization of the large-basis Fock matrix, $F[P]$.

Such a choice satisfies the four criteria above, by definition. Additionally, the special cases unique to dual-basis calculations are properly handled. For linear dependence among functions residing in the small basis only, the delta function projector preserves the mixing already used to account for these dependencies in the small-basis SCF, while producing properly orthonormal occupied MO coefficients. The reference energy in a dual-basis calculation is thus unambiguously defined. For linear dependence among large-basis functions, canonical orthogonalization performs adequately ($X^2S = 1$) but only for *exact* linear dependence. The delta function projector produces null large-basis elements of the occupied MO coefficients for this case, even for the more likely event of purely numerical linear dependence. Finally, for the more complicated case of a new, large-basis function becoming linearly dependent with a small-basis function, we have some choice in the projection matrix. Items 1 and 2 above impose two standard constraints on our projection, whereas items 3 and 4 impose constraints specific to dual-basis calculations, uniquely satisfied by the delta function projector. This choice of projection also avoids unnecessary complications in the analytic gradient of the dual-basis energy.

While the reference energy is uniquely defined by T, the dual-basis energy correction is still dependent on the choice of X. For the H-atom "thought experiment" above, the dual-basis correction will be nonzero (and dependent on the X used in the orthogonalized Roothaan equations, for example). However, this situation is identical to a standard SCF calculation. Differences among orthogonalization routines will be minimal, and the widespread use of canonical orthogonalization should further minimize these small discrepancies.

This delta function projector is easily implemented. Since the AO ordering is not {small + large} but instead is atom-ordered, a loop over elements of $S_{\mu\bar{\nu}}$ is required to distinguish between small and large basis functions. A value of 1.0 (to within the job-specified precision) in a column of this rectangular overlap matrix designates the column-index basis function as a small-basis function. The remaining functions are large-basis functions and correspond to null columns in the projection matrix. The same projection scheme is used for projections within the dual-basis analytic gradient.

One final caveat remains. The delta function projector is the method of choice for subset constructions, as defined in this paper. For 6-31G* ← 6-4G calculations,[67] in which the small basis is a subset by *primitives* only, the original projection scheme must be utilized. In practice, this fact is not a limitation, as linear dependencies most likely will never be problematic for these small basis sets. A case-dependent switch in the projection code allows for either type to be handled.

References

(1) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

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

(3) Becke, A. D. *J. Chem. Phys.* **1992**, *96*, 2155–2160.

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

(5) Andzelm, J.; Wimmer, E. *J. Chem. Phys.* **1992**, *96*, 1280–1303.

(6) Johnson, B. G.; Gill, P. M. W.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 5612–5626.

(7) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524–4529.

(8) Delley, B. *J. Chem. Phys.* **1991**, *94*, 7245–7250.

(9) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.

(10) Kristyàn, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

(11) Langreth, D. C.; Perdew, J. P. *Phys. Rev. B* **1980**, *21*, 5469–5493.

(12) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1986**, *33*, 8800–8802.

(13) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(14) Perdew, J. P. *Phys. Rev. B* **1988**, *34*, 7406.

(15) Langreth, D. C.; Mehl, M. J. *Phys. Rev. B* **1983**, *28*, 1809–1834.

(16) Langreth, D. C.; Mehl, M. J. *Phys. Rev. B* **1983**, *29*, 2310.

(17) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

(18) Jureka, P.; erný, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2006**, *28*, 555–569.

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

(20) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(21) Thonhauser, T.; Puzder, A.; Langreth, D. C. *J. Chem. Phys.* **2006**, *124*, 164106.

(22) Thonhauser, T.; Cooper, V. R. S; Li, A. P.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. Lett.* **2007**, *76*, 125112.

(23) Puzder, A.; Dion, M.; Langreth, D. C. *J. Chem. Phys.* **2006**, *124*, 164105.

(24) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(25) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

(26) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

(27) Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301–306.

(28) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201–033204.

(29) Heβelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

(30) Heβelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

(31) Heβelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

(32) Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646–659.

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

(34) DiStasio, R. A.; von Helden, G.; Steele, R. P.; Head-Gordon, M. *Chem. Phys. Lett.* **2007**, *437*, 277–283.

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

(36) Park, Y. C.; Lee, J. S. *J. Phys. Chem. A* **2006**, *110*, 5091–5095.

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

(38) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

(39) Takatani, T.; Sherrill, C. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6106–6114.

(40) DiStasio, R. A., Jr.; Steele, R. P.; Rhee, Y. M.; Shao, Y.; Head-Gordon, M. *J. Comput. Chem.* **2007**, *28*, 839.

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

(42) Eichkorn, K.; Treutler, O.; Ohm, H.; Heiser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.

(43) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.

(44) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359.

(45) Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6692.

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

(47) Antony, J.; Grimme, S. *J. Phys. Chem. A* **2007**, *111*, 4862–4868.

(48) Lochan, R. C.; Shao, Y.; Head-Gordon, M. *J. Chem. Theory Comput.* **2007**, *3*, 988–1003.

(49) Jung, Y.; Shao, Y.; Head-Gordon, M. *J. Comput. Chem.* **2007**, *28*, 1953–1964.

(50) Lochan, R. C.; Head-Gordon, M. *J. Chem. Phys.* **2007**, *126*, 164101.

(51) Rhee, Y. M.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 5314–5326.

(52) DiStasio, R. A., Jr.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073–1083.

(53) Hill, J. G.; Platts, J. A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2785–2791.

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

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

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

(57) 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.

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

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

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

(61) Liang, W.-Z.; Head-Gordon, M. *J. Phys. Chem. A* **2004**, *108*, 3206–3210.

(62) Steele, R. P.; DiStasio, R. A., Jr.; Shao, Y.; Kong, J.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074108.

(63) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2003**, *118*, 9497–9503.

(64) Steele, R. P.; Shao, Y.; DiStasio, R. A.; Head-Gordon, M. *J. Phys. Chem. A* **2006**, *110*, 13915–13922.

(65) DiStasio, R. A., Jr.; Steele, R. P.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 2731–2742.

(66) Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2006**, *124*, 184108.

(67) Steele, R. P.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 2455–2473.

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

(69) Kendall, R. A.; Dunning, T. H., Jr. *Chem. Phys. Lett.* **1992**, *96*, 6796.

(70) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.

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

(72) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104–6114.

(73) Armour, E. A. G.; Franz, J.; Tennyson, J. *Explicitly Correlated Wavefunctions*, 2nd ed.; Collaborative Computational Project on Molecular Quantum Dynamics, Daresbury Laboratory, Daresbury, Warrington, U.K., 2006.

(74) Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 4607.

(75) May, A. J.; Manby, F. R. *J. Chem. Phys.* **2004**, *121*, 4479.

(76) Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2006**, *124*, 054114.

(77) Manby, F. R.; Werner, H.-J.; Adler, T. B.; May, A. J. *J. Chem. Phys.* **2006**, *124*, 094103.

(78) Klopper, W.; Samson, C. C. M. *J. Chem. Phys.* **2002**, *116*, 6397.

(79) Kutzelnigg, W.; Klopper, W. *J. Chem. Phys.* **1991**, *94*, 1985.

(80) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190.

(81) Jureka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(82) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221–7230.

(83) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764–7776.

(84) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374–7383.

(85) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393.

(86) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.

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

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

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

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

(91) Mintz, B.; Lennox, K. P.; Wilson, A. K. *J. Chem. Phys.* **2004**, *121*, 5629–5634.

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

(93) Pitoák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *Chem. Phys. Chem.* **2008**, *9*, 1636–1644.

(94) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(95) Handy, N. C.; Schaefer, H. F., III *J. Chem. Phys.* **1984**, *81*, 5031.

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

(97) Byrd, E.; Sherrill, C.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 9736–9747.

(98) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

(99) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.

(100) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.

(101) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

(102) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.

(103) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.

(104) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Publications, Inc.: Mineola, NY, 1982; pp 142−145.

(105) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++: The Art of Scientific Computing*; Oxford University Press: Oxford, U.K., 1994; pp 62−68.

CT900058P

JCTC Journal of Chemical Theory and Computation

Electrostatically Embedded Many-Body Approximation for Systems of Water, Ammonia, and Sulfuric Acid and the Dependence of Its Performance on Embedding Charges

Hannah R. Leverentz and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received February 24, 2009

Abstract: This work tests the capability of the electrostatically embedded many-body (EE-MB) method to calculate accurate (relative to conventional calculations carried out at the same level of electronic structure theory and with the same basis set) binding energies of mixed clusters (as large as 9-mers) consisting of water, ammonia, sulfuric acid, and ammonium and bisulfate ions. This work also investigates the dependence of the accuracy of the EE-MB approximation on the type and origin of the charges used for electrostatically embedding these clusters. The conclusions reached are that for all of the clusters and sets of embedding charges studied in this work, the electrostatically embedded three-body (EE-3B) approximation is capable of consistently yielding relative errors of less than 1% and an average relative absolute error of only 0.3%, and that the performance of the EE-MB approximation does not depend strongly on the specific set of embedding charges used. The electrostatically embedded pairwise approximation has errors about an order of magnitude larger than EE-3B. This study also explores the question of why the accuracy of the EE-MB approximation shows such little dependence on the types of embedding charges employed.

1. Introduction

To compute properties of a chemical system often requires one to find a balance between computational cost and accuracy. A variety of relatively low-cost classical mechanical and semiempirical quantum mechanical methods allow one to calculate the properties of large (hundreds to thousands of atoms) systems quickly (sometimes within a fraction of a second), but, without problematic parametrization against experimental data, these methods are often incapable of providing more than qualitative accuracy for properties derived from a potential energy surface (PES). At the other extreme, calculations based on the first principles of quantum mechanics [such as coupled cluster[1] (CC) or configuration interaction[2] (CI) theory] have been developed that in principle could be carried to nearly arbitrary levels of quantitative accuracy[3] but that in practice may be used to calculate the energies only of systems containing a few atoms because of the methods' high computational cost. Thus, much effort has been expended in order to find a broadly applicable method that can accurately calculate the energy of a large system at a cost that would be reasonable for use in either molecular dynamics (MD) or Monte Carlo (MC) simulations. Fragment-based approaches[4−13] are one class of methods that attempt to accomplish this goal. These methods involve breaking the large system into subsystems (which will be called fragments) that are small enough to be treated at some desired level of electronic structure theory. Often, an attempt is made to polarize each fragment by representing the "missing" fragments as point charges or continuous charge density distributions, and the large system's total energy is then calculated as some linear combination of the fragments' energies and sometimes of the energies of pairs and trimers of the fragments as well.

The electrostatically embedded many-body (EE-MB) method,[13−17] which will be described in greater detail in Section 2, is a relatively simple fragment-based method that is computationally inexpensive because it does not involve

10.1021/ct900095d CCC: $40.75
Published on Web 05/01/2009

the self-consistent determination of embedding point charges or charge distributions. In the formal EE-MB approximation, each fragment (or monomer), pair of fragments (dimer), and sometimes group of three or more fragments (trimer or higher oligomer), is embedded in a predetermined set of point charges (called embedding charges or background charges) that represents the fragments that are not explicitly included in the electronic structure calculation of a given monomer, dimer, or trimer. When tested on water clusters and on mixed clusters of water and ammonia, the EE-MB approximation showed itself to be a cost-effective way to accurately calculate the total energy of a system of noncovalently interacting molecules at virtually any desired level of electronic structure theory.[13−17] The present work continues to explore the EE-MB approximation by looking at two additional aspects of the EE-MB calculations, as described in the next two paragraphs.

First, the present study applies the EE-MB approximation to more complicated mixed systems than any on which it has yet been tested; the largest clusters considered in this article are formed from six water molecules, one ammonia molecule, and two sulfuric acid molecules. Clusters of this type were selected because these molecules are thought to be the fundamental components of clusters formed during the early stages of atmospheric nucleation processes.[18] In addition, these clusters test the EE-MB approximation's ability to predict accurate energies (compared to the "full" quantum mechanical calculation by the same electronic structure method) for systems involving both large and small fragments (the large fragment being sulfuric acid with five heavy atoms and the small fragments being water and ammonia with only one heavy atom each) as well as ions or charge transfer complexes because several of the configurations considered in this article correspond to clusters of ammonia, sulfuric acid, bisulfate ion, ammonium ion, and water rather than clusters of only ammonia, sulfuric acid, and water.

Second, the present study compares various ways to obtain the embedding charges and tests how sensitively the accuracy of the EE-MB approximation depends on the resulting sets of embedding charges. Typically the sets of background charges that represent the "missing" monomers are determined by performing some kind of population analysis or charge analysis on the electron density matrices of the isolated and optimized gas-phase monomers. Using these predetermined sets of background charges has several advantages relative to using charges that depend on the configuration under consideration: (1) it lowers the cost of the EE-MB calculation by precluding the need to perform additional self-consistent field calculations to determine the "best" background charges for each configuration, and (2) it maintains the straightforward availability of analytic gradients and Hessians (if they are already available for a given method of electronic structure theory) by removing the embedding charges' dependence on the specific geometry of the system. However, one might argue that using such an inflexible set of embedding charges may not adequately polarize each fragment and could potentially compromise the accuracy of the EE-MB approximation. Therefore, in the present study we also test some inexpensive ways to obtain embedding charges that *do* depend on the specific geometry of each system being studied, and we compare the EE-MB results from those geometry-dependent (GD) charges with those from the geometry-independent (GI) charges that would be used in the formal EE-MB approximation. One should note that the formal EE-MB approximation would be more easily applied to dynamical simulations[17] that require fast calculations of PES gradients, but that either the formal EE-MB approximation or one that uses geometry-dependent background charges would be convenient for Monte Carlo simulations, where the calculation of PES gradients is not required.

The outline of the rest of this paper is as follows: Section 2 briefly reviews the theoretical underpinnings of the EE-MB approximation, Section 3 describes the computational methods used to perform the tests in this study and also gives the details of how the various sets of background charges were obtained, Section 4 presents the results and discusses their significance, and Section 5 summarizes our conclusions.

2. Theory

The EE-MB approximation, like several other fragment-based methods, is based on the many-body expansion of a system's total energy. Once a system has been fragmented into N monomers, the many-body expansion expresses the system's total energy as a sum of the energetic contributions of the one-body (i.e., individual monomer) interactions (V_1), the two-body interactions (V_2), the three-body interactions (V_3), and so on up to the N-body term, as shown in eq 1.

$$E = V_1 + V_2 + V_3 + ... + V_N \tag{1}$$

If one denotes the energy of one of the monomers as though it had the geometry it has in the cluster but were alone in a vacuum as E_i (where i runs over the arbitrary labels given to the monomers), the energy of dimer as E_{ij}, and the energy of a trimer as E_{ijk}, then the first three terms on the right-hand side of eq 1 are defined in eqs 2 through 4; the definitions of the remaining terms can be inferred from these equations.

$$V_1 = \sum_{i=1}^{N} E_i \tag{2}$$

$$V_2 = \sum_{i<j}^{N} (E_{ij} - E_i - E_j) \tag{3}$$

$$V_3 = \sum_{i<j<k}^{N} [E_{ijk} - (E_{ij} - E_i - E_j) - (E_{ik} - E_i - E_k) - (E_{jk} - E_j - E_k) - E_i - E_j - E_k] \tag{4}$$

One could approximate the total energy of the system by truncating eq 1 at some term V_M with M less than N; this is the many-body (MB) approximation of the system's energy given by

$$E \approx V_1 + V_2 + V_3 + ... + V_M \tag{5}$$

If one truncates eq 1 after $M = 2$, one has made the two-body (2B) or pairwise additive (PA) approximation. If one

truncates eq 1 after $M = 3$, one has made the three-body (3B) approximation.

The same equations as above underlie the EE-MB approximation, but the EE-MB approximation accounts for some of the higher-body interactions in the lower-order terms by calculating the monomer, dimer, trimer, etc. energies (E_i, E_{ij}, and E_{ijk}) as though each monomer, dimer, trimer, etc. were embedded in a field of point charges located at the coordinates of the missing nuclei. The surrounding point charges polarize or distort the electronic orbitals of each monomer (or group of monomers) so that they take on shapes and amplitudes that more closely resemble those that they might have in the overall system's wave function or electron density. Some of the specific methods by which such sets of embedding charges could be obtained are described in Section 3.

3. Methods

3.1. Choices of Embedding Point Charges. The paper that introduced the EE-MB approximation[13] pointed out that there are two major categories by which background charges may be determined for use in an EE-MB calculation: The first category, which yields what in this work we call the geometry-dependent or GD charges, calculates the density matrix corresponding to the wave function or the electron density function of the entire system at a computationally inexpensive level of electronic structure theory, such as the semiempirical method AM1,[19] and performs a charge analysis (such as a Mulliken,[20] Löwdin,[21] or redistributed Löwdin[22] analysis) on that density matrix to calculate partial charges located at the system's atomic centers. The second category, which yields what in this work we call the geometry-independent or GI charges, calculates the optimized density matrix of each type of monomer involved in the system and performs a charge analysis on each of those density matrices. The individual monomers can have their density matrices optimized in either the gas phase or a liquid solution phase. For example, if the system being studied were a cluster of water molecules, one could optimize a water molecule as though it were isolated in the gas phase or one could use an implicit solvation model to mimic an aqueous solution around the water molecule. The atom-centered partial charges calculated from each individually optimized density matrix are then used as the point charges representing that type of monomer in the EE-MB calculation, regardless of that monomer's position or shape in the overall system. The original (i.e., formal) EE-MB approximation calculates GI charges from monomers optimized in the gas phase, but we test the following three general types of point charges in the present work: GD charges, GI charges from monomers optimized in the gas phase, and GI charges from monomers optimized in a solution phase. (One could imagine another type of GD charge where charges are calculated for monomers but at the geometry they have in the particular configuration of the whole system that is under consideration, but we will not consider this method).

3.2. Computational Methods. All EE-MB calculations carried out in the present work were conducted using the M06-2X[23] density functional, which was chosen because it performs better than other density functionals for noncovalent interactions between molecules composed of main-group elements.[23] Three different basis sets were used to test the overall accuracy of the EE-PA and the EE-3B approximation: MG3S,[24] cc-pV(T+d)Z+,[25] and aug-cc-pV(T+d)Z.[26]

In order to calculate the geometry-dependent (GD) sets of background charges, the AM1 wave function of each configuration studied was calculated, and the following methods of charge analysis were used on those wave functions: Mulliken population analysis,[20] Charge Model 1 (CM1A, where the A indicates that the version of CM1 used was specifically parametrized to be used with AM1 wave functions),[27] Charge Model 2 (CM2),[28] and Charge Model 3 (CM3).[29,30] One should note that Mulliken charges are Class II charges[27] and at best give electrostatic properties corresponding to an approximate level of theory with a finite set of basis functions, whereas CMx (x = 1A, 1P, 2, 3, 4, or 4M) charges are Class IV charges because they include empirical parameters that map Class II charges (such as Mulliken or Löwdin charges) to charges that more realistically reproduce experimental dipole moments. Following the recommendation of Udier-Blagovicì et al.,[31] a final set of GD charges has also been tested: these charges are simply CM1A charges scaled by 1.14, and they are labeled "CM1A*1.14" or "scaled CM1A" charges. The scaling is designed to make the charges (although computed in the gas-phase) more appropriate for liquid simulations.

Geometry-independent (GI) charges were obtained from the density matrices of both gas-phase and liquid-phase monomers. The optimized density matrices of gas-phase monomers were used to calculate point charges according to eight different methods of charge analysis: ChElPG,[32] Merz−Singh−Kollman (MK),[33,34] the MK method with the additional constraint to reproduce dipole moments as well as electrostatic potentials (ESP-Dipole; see the Gaussian 03 online manual[35] for details), Natural Bond Orbital (NBO),[36] CM1A, CM2, CM3, and CM4M.[37,38] NBO can be considered to be a Class II charge model, but rather than calculating charges from a density matrix expressed in terms of the original basis set functions, NBO charges are calculated from a density matrix expressed in terms of a set of functions that adopt the "natural" shapes that a chemist would expect to describe various types of chemical bonds. The ChElPG, ESP-Dipole, and MK methods are quite similar to one another and yield what are classified as Class III charges; these charges are those that best reproduce the electrostatic potential due to a system's electron density distribution function at various points in space around the system (which is a gas-phase monomer for GI charge analysis). The ChElPG, ESP-Dipole, MK, and NBO charges were determined from electron density functions computed by M06-2X/cc-pV(T+d)Z+//M06-2X/cc-pV(T+d)Z+ (we adopt the common notation *W/X//Y/Z*, where *Y* is the level of electronic structure theory or density functional and *Z* is the basis set with which the geometry of the system was optimized, and where *W* is the level of electronic structure theory or density functional and *X* is the basis set with which the electron density and/or energy to be used in subsequent calculations

was optimized). Because CM4M contains parameters that depend on the density functional and basis set chosen and is currently parametrized for a variety of double-ζ but not triple-ζ quality basis sets, the CM4M charges were calculated from the M06-2X/MIDI!//M06-2X/cc-pV(T+d)Z+ density matrix (the MIDI![39] basis set is of double-ζ quality and was designed specifically for the efficient calculation of accurate geometries and partial charges). Charge Models 1, 2, and 3 were originally parametrized for semiempirical methods and to obtain the geometry-independent CM1A, CM2, and CM3 charges we used the AM1//AM1 wave functions of the isolated gas-phase monomers. The scaled CM1A charges (CM1A*1.14) of the gas-phase monomers were also used as GI background charges for EE-MB calculations.

Three sets of background charges were based on liquid-phase monomers: these charges are denoted SM5.42/CM2, SM8/CM4M, and SMD/CM4M. To describe the density matrices used to obtain the charges, we adopt the following notation: *slvnt-SMx/W/X//slvnt-SMy/Y/Z*, where *W*, *X*, *Y*, and *Z* are as defined above and where *SMy* is the solvation model applied to perform a liquid-phase geometry optimization of the monomer, where *SMx* is the solvation model used to obtain a liquid-phase optimized wave (or density) function for subsequent charge analysis (for this study, *x* and *y* can be 5.42, 8, or D) and where *slvnt* indicates the solvent in which the monomer was theoretically immersed (for this study, *slvnt* = aq to signify that the calculation was performed in an aqueous solution). For SM5.42/CM2 charges, CM2 charges were calculated from the aq-SM5.42/AM1//AM1 monomer density matrices; that is, the monomer geometries were optimized by AM1 in the gas phase, and the wave functions were then optimized by AM1 in the aqueous phase using Solvation Model 5.42. "SM8" and "SMD" indicate Solvation Model 8[40] and Solvation Model D,[41] respectively. The SM8/CM4M charges are CM4M charges calculated from the aq-SM8/M06-L/6-31G(d)//M06-L/6-31G(d) monomer density matrices based on the M06-L density functional[42] with the 6-31G(d) basis,[43,44] and the SMD/CM4M charges are CM4M charges calculated from the aq-SMD/M06-L/6-31G(d)//M06-L/6-31G(d) monomer density matrices.

All geometries were optimized using the Minnesota Gaussian Functional Module, version 3.0 (MN-GFM-v3.0),[45] a locally modified version of the GAUSSIAN 03[46] electronic structure package, revision D.01. MN-GFM-v3.0 was also used to perform the charge analyses for the ChElPG, ESP-Dipole, MK, and NBO charges and to carry out single-point energy calculations on the clusters and molecules involved in this study. The Minnesota Gaussian Solvation Module, version 2008 (MN-GSM-v.2008),[47] a module for performing solvation calculations in GAUSSIAN 03, revision D.01, was used to compute the CM4M and SM8/CM4M charges. The SMD/M06-L/6-31G(d)//M06-L/6-31G(d) wave function was computed using the GESOL[48] program (an external module for GAUSSIAN 03), but the SMD/CM4M charges based on this wave function were computed using MN-GSM-v.2008. Calculations done on AM1 wave functions to find the geometry-dependent and geometry-independent CM1A, CM2, and CM3 charges as well as the geometry-independent SM5.42/CM2 charges were performed using AMSOL-version 7.1.[49] All EE-MB calculations were executed using MBPAC 2007-2,[50] a program that calls GAUSSIAN 03 or MN-GFM to perform electrostatically embedded many-body calculations.

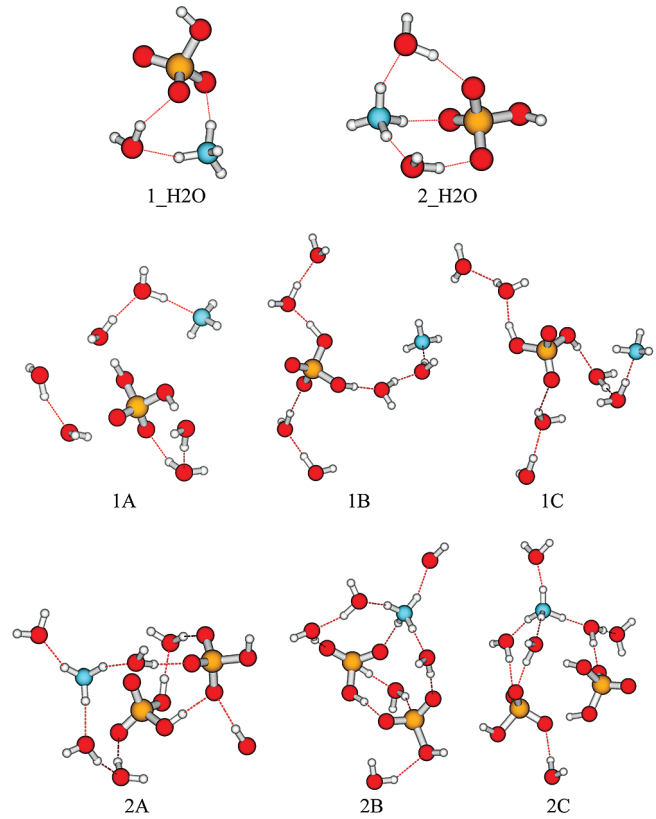


Figure 1. Eight clusters formed from water, ammonia, and sulfuric acid (note that many of the clusters contain ammonium and bisulfate ions rather than neutral ammonia and sulfuric acid molecules). The 1_H2O and 2_H2O structures were optimized at M06-2X/MG3S. The remaining structures were taken from molecular dynamics simulations. The composition of each cluster is as follows: (a) 1_H2O = $(HSO_4^-)(NH_4^+)(H_2O)$, (b) 2_H2O = $(HSO_4^-)(NH_4^+)(H_2O)_2$, (c) 1A = 1B = 1C = $(H_2SO_4)(NH_3)(H_2O)_6$, and (d) 2A = 2B = 2C = $(H_2SO_4)\ (HSO_4^-)(NH_4^+)(H_2O)_6$.

4. Results and Discussion

The first goal of this study was simply to test how well the EE-PA and EE-3B approximations are able to reproduce the energies of systems containing water, ammonia, and sulfuric acid and/or their conjugate acids or bases calculated in the conventional manner for a given model chemistry (for this discussion, "model chemistry" or "method" implies a specific combination of electronic structure theory level or density functional with a specific basis set). To do this, the binding energies of eight clusters, each composed of between three and nine molecules, were calculated by M06-2X with three different basis sets: MG3S, cc-pV(T+d)Z+, and aug-cc-pV(T+d)Z.

The eight clusters considered in the first part of this work and their names are shown in Figure 1. The smallest cluster, called 1_H2O, contains one bisulfate ion, one ammonium ion, and only one water molecule. Because the 1_H2O cluster

Table 1. Binding Energies (E_{bind}, kcal/mol) of Eight Clusters with M06-2X Density Functional and Three Basis Sets and the Corresponding Errors[a] (kcal/mol) from the EE-PA and EE-3B Calculations using Five Different Sets of Geometry-Independent Background Charges

			EE-PA errors					EE-3B errors				
basis	system	E_{bind} full	ChElPG	ESP-dipole	MK	NBO	CM4M	ChElPG	ESP-dipole	MK	NBO	CM4M
MG3S	1_H2O	−30.50	−0.57	−0.62	−0.59	−0.12	−0.98					
cc-pV(T+d)Z+	1_H2O	−29.69	−0.61	−0.65	−0.63	−0.17	−1.04					
aug-cc-pV(T+d)Z	1_H2O	−29.57	−0.34	−0.38	−0.36	−0.06	−0.66					
MG3S	2_H2O	−46.14	−1.36	−1.42	−1.36	−0.74	−1.87	0.06	0.06	0.06	0.02	0.06
cc-pV(T+d)Z+	2_H2O	−45.31	−1.43	−1.48	−1.43	−0.82	−1.96	0.03	0.04	0.03	−0.01	0.04
aug-cc-pV(T+d)Z	2_H2O	−45.08	−1.16	−1.20	−1.16	−0.65	−1.56	0.03	0.04	0.03	0.00	0.01
MG3S	1A	−52.23	1.11	1.17	1.06	0.04	2.19	0.17	0.17	0.17	0.11	0.22
cc-pV(T+d)Z+	1A	−51.80	1.14	1.18	1.08	0.01	2.23	0.15	0.15	0.14	0.10	0.19
aug-cc-pV(T+d)Z	1A	−50.87	1.23	1.31	1.17	−0.06	2.18	0.16	0.16	0.15	0.08	0.19
MG3S	1B	−31.39	0.48	0.50	0.44	−0.26	1.31	−0.06	−0.07	−0.06	−0.06	−0.05
cc-pV(T+d)Z+	1B	−31.80	0.50	0.51	0.45	−0.27	1.32	−0.05	−0.05	−0.05	−0.05	−0.04
aug-cc-pV(T+d)Z	1B	−30.89	0.53	0.58	0.48	−0.39	1.27	−0.04	−0.04	−0.04	−0.03	−0.03
MG3S	1C	−43.89	0.61	0.63	0.55	0.07	1.29	0.14	0.14	0.13	0.10	0.17
cc-pV(T+d)Z+	1C	−43.36	0.68	0.70	0.62	0.09	1.37	0.15	0.15	0.15	0.12	0.18
aug-cc-pV(T+d)Z	1C	−42.31	0.79	0.84	0.73	0.06	1.38	0.13	0.13	0.13	0.11	0.15
MG3S	2A	−66.84	−2.18	−2.23	−2.14	−1.58	−2.83	0.11	0.11	0.10	0.10	0.12
cc-pV(T+d)Z+	2A	−66.30	−2.21	−2.25	−2.16	−1.67	−2.82	0.14	0.13	0.13	0.12	0.15
aug-cc-pV(T+d)Z	2A	−65.24	−2.24	−2.28	−2.20	−1.76	−2.73	0.22	0.21	0.21	0.06	0.26
MG3S	2B	−68.53	−2.54	−2.61	−2.50	−1.79	−3.18	0.12	0.13	0.12	0.06	0.17
cc-pV(T+d)Z+	2B	−67.65	−2.63	−2.68	−2.58	−1.91	−3.23	0.14	0.14	0.13	0.07	0.20
aug-cc-pV(T+d)Z	2B	−66.71	−2.56	−2.60	−2.51	−1.83	−3.01	0.22	0.23	0.22	0.06	0.30
MG3S	2C	−62.70	−4.00	−4.15	−3.97	−2.55	−5.28	0.34	0.34	0.33	0.31	0.37
cc-pV(T+d)Z+	2C	−62.13	−4.03	−4.17	−3.99	−2.65	−5.27	0.29	0.29	0.27	0.26	0.30
aug-cc-pV(T+d)Z	2C	−61.29	−4.07	−4.22	−4.04	−2.58	−5.12	0.38	0.38	0.37	0.27	0.41

[a] The errors are calculated as E_{bind}(EE-MB) − E_{bind}(full).

contains only three molecules and because each molecule is defined as a monomer for the EE-MB calculations done in this study, only the EE-PA approximation is applied to this cluster (the EE-3B approximation necessarily yields the same result as the conventional calculation for the same method). The next smallest cluster contains one bisulfate ion, one ammonium ion, and two water molecules and is called the 2_H2O cluster. Clusters 1_H2O and 2_H2O are based on structures shown in Figure 1 of ref 51; however, the precise coordinates used for the single-point energy calculations carried out in this study (which have been included in Supporting Information) were the result of M06-2X/MG3S geometry optimizations done on these two clusters. Six clusters, three of which comprise six water molecules, one ammonia molecule, and one sulfuric acid molecule and three of which comprise six water−water molecules, one ammonium ion, one bisulfate ion, and one sulfuric acid molecule, were also studied. These configurations are called 1A, 1B, 1C, 2A, 2B, and 2C. They were generated during an MD simulation;[52] a "1" in the name indicates that one sulfuric acid molecule was used in the starting configuration of the simulation, and a "2" indicates that two sulfuric acid molecules (one of which is in bisulfate form) were used in the starting configuration of the simulation.

The binding energy (E_{bind}) of each of the clusters described in the preceding paragraph was first calculated in the conventional way with respect to the neutral gas-phase monomers with geometries optimized by the M06-2X/MG3S method. That is,

$$E_{bind} = E_{cluster} - \sum E_{molecule} \quad (6)$$

where $E_{cluster}$ is the M06-2X/basis//M06-2X/MG3S absolute electronic energy of the cluster and $E_{molecule}$ is the M06-2X/basis//M06-2X/MG3S absolute energy of the neutral version of each molecule from which the cluster is formed [basis = MG3S, cc-pV(T+d)Z+, or aug-cc-pV(T+d)Z]. Table 1 lists the binding energy of each cluster when calculated in the conventional (or "full") manner by each method and the difference (or error) between the EE-PA calculation with different sets of geometry-independent background charges calculated according to the formal EE-MB prescription; that is, the background charges were calculated from the optimized gas-phase monomers. These full binding energies were then used to test the truncated EE-MB expansions. Note that although the binding energies are defined with respect to the neutral versions of the monomers, in applying the EE versions of eqs 1−4 the clusters were fragmented into both neutral and ionic monomers and the background charges that were used to represent the missing monomers were calculated from the corresponding optimized gas-phase monomers, which are both neutral and ionic.

Table 1 shows that the EE-MB approximation, and in particular the EE-3B approximation, continues to yield accurate results when compared to the conventional calculations when it is used to calculate the binding energies of these more complicated clusters than others on which it has so far been tested. The maximum absolute error for the EE-PA calculations is 5.28 kcal/mol; this error occurs for structure 2C, which has a large binding energy, and so it corresponds to an error of only 8.4%. Furthermore, the average relative absolute error over all eight configurations, three basis sets, and five background charge sets for the EE-PA approximation is only 3.0%. The maximum absolute error for the EE-3B calculations (seven configurations) is a mere 0.41 kcal/mol, again for structure 2C, and it corresponds to a relative absolute error of only 0.7%. The average relative

absolute error over all seven configurations, three basis sets, and five background charge sets for the EE-3B approximation is 0.3%. These results imply that the EE-3B approximation is capable of handling complicated systems involving ions and/or charge transfer complexes and a wide range of monomer sizes and monomer complexity.

One particularly striking aspect of Table 1 is the comparative size of the EE-3B error in reproducing the full calculations and the deviations of the various full calculations from one another. Although all three basis sets are multiply polarized valence triple-ζ sets with diffuse functions, the results for a given cluster with a pair of basis sets differ from one another on average by 0.9 kcal/mol whereas a typical error due to the EE-3B approximation is ~0.1 kcal/mol. Thus the error incurred by truncating the EE-MB expansion with geometry-independent background charges is much less than the uncertainty due to choice of basis set.

The second goal of this study was to compare two major categories of methods by which background charges for EE-MB calculations can be obtained: geometry-dependent and geometry-independent charges. From the geometry-independent charges, two subcategories of partial charge calculation methods are also compared: those that use the gas-phase monomer density matrices and those that use liquid-phase monomer density matrices. (See Section 3.1 for a more detailed explanation of these categories.) This portion of the study focuses on the binding energies of the three largest (and most complex) clusters (2A, 2B, and 2C) calculated by the M06-2X/cc-pV(T+d)Z+ method both conventionally and with the EE-MB approximation (MB = PA or 3B for this study) using background charge sets from each of the above categories and subcategories. [The cc-pV(T+d)Z+ basis set was selected because, of the three basis sets shown in Table 1, it is generally the most efficient at reducing basis set superposition error (BSSE); that is, on average the cc-pV(T+d)Z+ basis set yields about the same amount of BSSE as aug-cc-pV(T+d)Z but at lower cost.[25] The MG3S basis set tends to yield larger amounts of BSSE than the other two basis sets. Using a basis set that in general yields low BSSE diminishes the need to attempt to correct for BSSE by methods such as counterpoise correction,[53] which would significantly increase the overall cost of a calculation of the binding energy of a nine-molecule system.[54]]

Table 2 lists the errors from the EE-PA and EE-3B calculations relative to the conventionally calculated binding energy for each of the three clusters. The charge sets labeled with the prefix "GI_" are geometry-independent charge sets; that is, these charges were calculated from the density matrices of individual monomers and therefore do not depend on a given cluster's geometry (in still other words, these charge sets would remain the same for any cluster containing the same types of molecules and/or ions). Of the geometry-independent charge sets, those that were computed from aqueous-phase monomer density matrices contain the letters "SM" (for "solvation model") in their labels immediately following the "GI_" prefix; those that were computed from density matrices of gas-phase monomers do not. The "GD_" prefix indicates that the given charge set is geometry dependent; that is, the charge set is calculated from a semiempirical wave function for the entire cluster and therefore depends on the geometry of the cluster. Table 3 summarizes the results shown in Table 2 by listing the mean unsigned errors (MUE) and root mean squared errors (RMSE) of the EE-PA and EE-3B approximations over all three clusters. (Section 3.2 contains the specific description of the meaning of the name of each charge set and the method by which each charge set was calculated).

Table 2. Binding Energies [E_{bind}(full), kcal/mol] from Conventional Calculations at M06-2X/cc-pV(T+d)Z+ and the Corresponding Errors[a] (kcal/mol) from the EE-MB Calculations When Different Sets of Background Charges Are Used

	2A		2B		2C	
charge model	EE-PA	EE-3B	EE-PA	EE-3B	EE-PA	EE-3B
E_{bind}(full)	−66.30		−67.65		−62.13	
none[b]	−4.94	−0.85	−8.51	−0.14	−15.99	−1.11
GI_ChElPG[c]	−2.21	0.14	−2.63	0.14	−4.03	0.29
GI_ESP-Dipole	−2.25	0.13	−2.68	0.14	−4.17	0.29
GI_MK	−2.16	0.13	−2.58	0.13	−3.99	0.27
GI_NBO	−1.67	0.12	−1.91	0.07	−2.65	0.26
GI_CM4M	−2.82	0.15	−3.23	0.20	−5.27	0.30
GI_CM1A	−2.29	0.14	−2.71	0.15	−4.21	0.31
GI_(CM1A*1.14)	−2.37	0.21	−2.76	0.23	−3.75	0.37
GI_CM2	−2.33	0.19	−2.83	0.16	−4.06	0.44
GI_CM3	−2.42	0.18	−2.93	0.17	−4.31	0.43
GI_SM5.42/CM2	−2.10	0.17	−2.50	0.12	−3.53	0.39
GI_SM8/CM4M	−2.40	0.14	−2.76	0.16	−4.36	0.26
GI_SMD/CM4M	−2.35	0.14	−2.70	0.15	−4.24	0.26
GD_AM1-Mulliken[d]	−2.96	0.18	−3.82	0.20	−6.42	0.52
GD_CM1A	−1.81	0.10	−2.04	0.08	−3.36	0.18
GD_CM1A*1.14	−1.98	0.15	−2.16	0.12	−2.85	0.21
GD_AM1-CM2	−1.79	0.11	−2.20	0.06	−3.67	0.29
GD_AM1-CM3	−1.80	0.11	−2.18	0.08	−3.50	0.26

[a] The errors are calculated as E_{bind}(EE-MB) − E_{bind}(full). [b] "None" implies that no electrostatic embedding was used for these calculations; i.e., this row gives PA and 3B errors, not EE-PA and EE-3B errors. [c] The "GI_" prefix indicates that these background charges are geometry independent. An "SM" following this prefix indicates that the charges were obtained from aqueous-phase monomers; all others were obtained from gas-phase monomers. [d] The "GD_" prefix indicates that these background charges are geometry dependent.

First, Tables 2 and 3 show that electrostatic embedding significantly enhances the accuracy of the PA and 3B approximations. Without electrostatic embedding, the pairwise additive MUE is 9.82 kcal/mol, whereas the maximum electrostatically embedded pairwise additive MUE is 4.40 kcal/mol and the average EE-PA MUE is 2.96 kcal/mol. Similarly, the three-body MUE without electrostatic embedding is 0.70 kcal/mol, whereas the maximum EE-3B MUE is 0.30 kcal/mol and the average EE-3B MUE is 0.20 kcal/mol.

A second point illustrated by Tables 2 and 3 is that GD charge sets yield EE-MB results that are only slightly better than those from GI sets. For the EE-PA approximation, the average GI MUE is 3.00 kcal/mol and the average GD MUE is 2.84 kcal/mol. For the EE-3B approximation, the average GI MUE is 0.21 kcal/mol and the average GD MUE is 0.18 kcal/mol. The small (nearly insignificant for the EE-3B approximation) improvement in accuracy afforded by the GD charge sets is not worth the loss of convenient analytic gradients when performing MD simulations, nor does it even seem to be worth the tiny relative increase in cost that would

Table 3. Mean Unsigned Errors (MUE) and Root Mean Squared Errors (RMSE) in kcal/mol over Three Configurations (2A, 2B, and 2C) of an $(H_2SO_4)(HSO_4^-)(NH_4^+)(H_2O)_6$ System[a]

	EE-PA		EE-3B	
charge model	MUE	RMSE	MUE	RMSE
full	0.00	0.00	0.00	0.00
none[b]	9.82	10.84	0.70	0.81
GI_ChElPG[c]	2.96	3.06	0.19	0.20
GI_ESP-Dipole	3.03	3.14	0.19	0.20
GI_MK	2.91	3.01	0.18	0.19
GI_NBO	2.08	2.12	0.15	0.17
GI_CM4M	3.77	3.92	0.22	0.23
GI_CM1A	3.07	3.18	0.20	0.22
GI_(CM1A*1.14)	2.96	3.02	0.27	0.28
GI_CM2	3.07	3.16	0.26	0.29
GI_CM3	3.22	3.32	0.26	0.29
GI_SM5.42/CM2	2.71	2.78	0.23	0.26
GI_SM8/CM4M	3.17	3.28	0.19	0.19
GI_SMD/CM4M	3.10	3.21	0.18	0.19
GD_AM1-Mulliken[d]	4.40	4.64	0.30	0.34
GD_CM1A	2.41	2.50	0.12	0.13
GD_CM1A*1.14	2.33	2.36	0.16	0.17
GD_AM1-CM2	2.56	2.68	0.15	0.18
GD_AM1-CM3	2.50	2.60	0.15	0.17

[a] The full (or conventional) and EE-MB calculations were performed by the M06-2X/cc-pV(T+d)Z+ method. [b] No background charges were used for these calculations; i.e., these are the PA and 3B approximations to the total energy without electrostatic embedding. [c] The "GI_" prefix indicates that these background charges are geometry independent. An "SM" following this prefix indicates that the charges were obtained from aqueous-phase monomers; all others were obtained from gas-phase monomers. [d] The "GD_" prefix indicates that these background charges are geometry dependent.

be incurred during MC simulations. Therefore, the original EE-MB approximation where the electrostatic embedding is based on GI charges continues to be the recommended approach for EE-MB calculations.

A third conclusion that may be drawn from Tables 2 and 3 is that using gas-phase monomer wave or density functions as a starting point for the determination of GI charges is just about as good as using charges derived from liquid-phase monomers. The average MUE of the charge sets obtained from gas-phase monomers is 3.01 kcal/mol for the EE-PA approximation and 0.21 kcal/mol for the EE-3B approximation. The average MUEs of the charge sets obtained from aqueous-phase monomers are 2.99 and 0.20 kcal/mol, respectively. This may come as a surprise because one might expect that the electron density around a monomer in a cluster would more closely resemble the electron density distribution of a solvated monomer than it would the electron density distribution of a gas-phase monomer. This is because a monomer in a cluster or a monomer in solution is polarized by the surrounding monomers and might experience more charge separation than would a monomer in the gas phase; i.e., charges derived from either a monomer in a cluster or a monomer in solution might take on more extreme magnitudes than charges derived from a monomer in the gas phase. However, this did not turn out to be the case. The charges derived from liquid-phase monomers were on the whole quite similar to those derived from gas-phase monomers, as shown in Tables 4−6. This explains why these charge sets produce similar results when used as the background charges in EE-MB calculations. Once again, the original EE-MB approximation (taking GI charges from monomers in the gas phase) remains the recommended approach because gas-phase monomer calculations are less costly (even if only by a little) than liquid-phase calculations and because potential ambiguity regarding which solvent to choose for monomers involved in mixed clusters is avoided when the gas-phase monomers are used to generate embedding charges.

Table 4. Geometry-Independent Background Charges (in *e*) Based on the Geometries of the Gas-Phase Monomers Optimized with the M06-2X/cc-pV(T+d)Z+ Method

molecule	atom type	ChElPG[a]	ESP-dipole[a]	MK[a]	NBO[a]	CM4M[b]
H_2O	O	−0.726	−0.709	−0.731	−0.930	−0.601
H_2O	H	0.363	0.355	0.365	0.465	0.300
HSO_4^-	S	1.428	1.329	1.328	2.591	0.403
HSO_4^-	O	−0.700	−0.677	−0.676	−1.016	−0.423
HSO_4^-	H	0.372	0.378	0.375	0.472	0.289
H_2SO_4	S	1.164	1.042	1.049	2.602	0.499
H_2SO_4	O	−0.509	−0.482	−0.483	−0.910	−0.298
H_2SO_4	H	0.435	0.443	0.442	0.519	0.347
NH_4^+	N	−0.784	−0.834	−0.834	−0.859	−0.604
NH_4^+	H	0.446	0.458	0.458	0.465	0.401

[a] Charge analyses were done on the M06-2X/cc-pV(T+d)Z+ gas-phase monomer density matrices. [b] Charge analyses were done on the M06-2X/MIDI! gas-phase monomer density matrices.

Table 5. Geometry-Independent Background Charges (in *e*) Based on the Geometries of the Gas-Phase Monomers Optimized with the AM1 Method

molecule	atom type	CM1A[a]	CM1A*1.14[a]	CM2[a]	CM3[a]	SM5.42/CM2[b]
H_2O	O	−0.706	−0.805	−0.711	−0.679	−0.783
H_2O	H	0.353	0.402	0.356	0.340	0.392
HSO_4^-	S	1.433	1.634	2.878	2.491	2.934
HSO_4^-	O	−0.709	−0.808	−1.061	−0.963	−1.086
HSO_4^-	H	0.401	0.457	0.368	0.360	0.409
H_2SO_4	S	1.440	1.642	2.874	2.487	2.961
H_2SO_4	O	−0.601	−0.685	−0.937	−0.842	−0.974
H_2SO_4	H	0.481	0.548	0.438	0.439	0.468
NH_4^+	N	−0.514	−0.586	−0.793	−0.829	−0.792
NH_4^+	H	0.378	0.431	0.448	0.457	0.448

[a] Charge analyses were done on the AM1 gas-phase monomer density matrices. [b] Charge analyses were done on the SM5.42/AM1 aqueous-phase monomer density matrices.

Table 6. Geometry-Independent Background Charges[a] (in *e*) Based on the Geometries of the Gas-Phase Monomers Optimized by the M06-L/6-31G(d) Method

molecule	atom + label	SM8/CM4M	SMD/CM4M
H_2O	O	−0.695	−0.708
H_2O	H	0.347	0.354
HSO_4^-	S	0.751	0.775
HSO_4^-	O	−0.525	−0.534
HSO_4^-	H	0.350	0.361
H_2SO_4	S	0.875	0.867
H_2SO_4	O	−0.416	−0.418
H_2SO_4	H	0.393	0.401
NH_4^+	N	−0.584	−0.592
NH_4^+	H	0.396	0.398

[a] Charge analyses were done on the SM*x*/M06 L/6-31G(d) (x = 8, D) aqueous-phase monomer density matrices.

Table 7. Dipoles[a] (in debye) of Individual Water Molecules within Different Configurations of an $(H_2SO_4)(HSO_4^-)(NH_4^+)(H_2O)_6$ System from Various Point Charge Representations of Those Configurations

cluster label	monomer label	Mulliken	CM1A	CM1A*1.14	CM2	CM3
2A	2	1.39	2.26	2.58	2.24	2.21
2A	4	1.37	2.20	2.50	2.18	2.15
2A	5	1.58	2.47	2.81	2.43	2.42
2A	6	1.35	2.30	2.62	2.31	2.24
2A	7	1.23	2.13	2.42	2.12	2.06
2A	9	1.30	2.22	2.53	2.21	2.15
2B	2	1.48	2.40	2.74	2.37	2.34
2B	4	1.39	2.24	2.55	2.22	2.19
2B	5	1.39	2.20	2.51	2.17	2.16
2B	6	1.30	2.19	2.49	2.20	2.13
2B	7	1.25	2.24	2.55	2.24	2.16
2B	9	1.31	2.19	2.50	2.18	2.13
2C	2	1.54	2.46	2.81	2.39	2.40
2C	4	1.42	2.25	2.56	2.21	2.20
2C	5	1.34	2.19	2.50	2.18	2.14
2C	6	1.44	2.44	2.79	2.45	2.38
2C	7	1.21	2.12	2.42	2.12	2.05
2C	9	1.18	2.08	2.38	2.10	2.01
average (debye)		1.36	2.25	2.57	2.24	2.20
standard deviation (debye)		0.11	0.12	0.13	0.11	0.12
% standard deviation		8.0	5.2	5.2	4.7	5.4

[a] Dipoles are calculated with respect to each water molecule's center of nuclear charge.

To summarize the discussion of the results presented so far, one could simply state that the accuracy of the EE-MB approximation does not appear to be heavily dependent on the set of background charges chosen. Compared to the average binding energy of the three clusters, −65 kcal/mol, a 5% error would be 3.3 kcal/mol and a 1% error would be 0.65 kcal/mol. Thus, most of the EE-PA calculations yield MUEs of less than 5% and all of the EE-3B calculations yield MUEs of less than 1%, regardless of the charge model used in this study.

The reason for which the two subcategories of GI charges (gas-phase vs liquid-phase monomers) do not produce significantly different EE-MB results was addressed in an earlier paragraph, but one is still left to wonder why GI charges manage to do about as well as GD charges. In an attempt to understand this, one can investigate (1) how the dipoles of individual water molecules vary within a cluster and (2) how the dipole of an individual water molecule varies when the water molecule is embedded in different sets of point charges. Because water is a planar molecule and generally possesses close to C_{2v} symmetry, its dipole is a good indicator of the extent to which the water molecule is polarized. (Note that the word "dipole" is used to mean "the magnitude of the dipole moment".) If the dipoles of the water molecules in the entire cluster do not vary much (Test 1), then one can see how the "inflexible" charges from a gas-phase monomer could adequately mimic the effects of other water molecules in the cluster. Additionally, if the dipole of a single water molecule embedded in point charges does not vary much with different background charge sets (Test 2), then one can infer that the choice of background charges will not strongly impact an EE-MB calculation, because the purpose of the background charges is to polarize the monomers, dimers, trimers, etc. If the embedding charges do not have a strong effect on the polarization of a monomer, then it is unlikely that they would have a strong effect on the result of an EE-MB calculation.

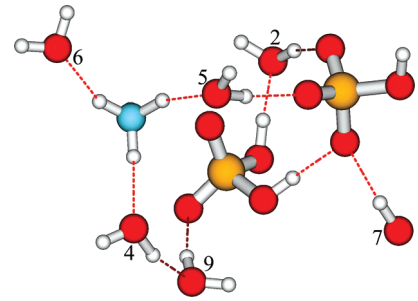


Figure 2. The 2A configuration with the water molecules labeled with the arbitrary fragment numbers that they were assigned for the EE-MB calculations. Note that two hydrogen atoms are obstructed from view in this figure: one is obstructed by the nitrogen atom of the ammonium ion, and the other is obstructed by the oxygen atom of water molecule 7.

As stated above, Test 1 investigates the variation in the dipoles of the water molecules within a given cluster. The dipole moment of each water molecule in clusters 2A, 2B, and 2C was calculated using the Mulliken, CM1A, scaled CM1A, CM2, and CM3 point charges that were obtained from the AM1 wave functions of those clusters; these dipoles are listed in Table 7. The point charges were used to calculate the dipoles because one cannot determine the expectation value of the dipole moment of an individual water molecule from the wave function of the entire cluster. The dipole moment of each water molecule was calculated with respect to that particular molecule's center of nuclear charge. To give an idea of the locations of the water molecules within a cluster, Figure 2 shows configuration 2A with each water molecule labeled by the arbitrary fragment number that it was assigned for the EE-MB calculations; these numbers correspond to the monomer labels listed in Table 7. The water molecules of configurations 2B and 2C were labeled in essentially the same way, although of course their locations within the cluster are slightly different in each case. Table 7 shows that for every type of point charge representation other

Table 8. Dipoles[a] (in debye) of Gas-Phase Monomers Calculated from Point Charges and Compared to Best Estimates

charge model	H_2O	H_2SO_4	NH_3	HSO_4^-	OH^-	MUE[b]	RMSE[b]
ChElPG	2.03	3.25	1.64	2.65	1.59	0.19	0.21
ESP-dipole	1.98	3.31	1.59	2.72	1.59	0.20	0.22
MK	2.04	3.27	1.67	2.68	1.59	0.21	0.22
NBO	2.60	3.70	1.89	3.24	2.20	0.68	0.70
CM4M	1.68	3.02	1.55	2.40	0.95	0.18	0.21
CM1A	2.02	2.25	1.76	1.83	1.38	0.40	0.49
CM1A*1.14	2.30	2.56	2.01	2.08	1.57	0.43	0.44
CM2	2.03	1.89	1.61	1.59	1.40	0.49	0.67
CM3	1.94	1.87	1.60	1.55	1.29	0.48	0.68
best estimate	1.85[c]	2.96[d]	1.47[e]	2.60[f]	1.33[f]		

[a] The dipole of each compound was calculated with respect to the compound's center of nuclear charge. [b] The MUEs/RMSEs are averages of the differences between the point-charge derived and the best estimate dipoles over all five compounds. The MUEs/RMSEs are given in debye. [c] References 55−57. [d] Reference 58. [e] Reference 59. [f] Determined in the present study by a finite-field calculation at CCSD(T)/aug-cc-pV(T+d)Z//M06-2X/cc-pV(T+d)Z+.

than Mulliken charges, the relative standard deviation is only 5%. In the case of dipoles derived from Mulliken charges the relative standard deviation is 8%. These results indicate that the variation in the extent of polarization of water molecules in a semiempirical calculation of an entire cluster's wave function is not large.

Test 2 investigates the dependence of the extent of polarization of a single embedded water molecule on the choice of background charges. Monomer number 5 of configuration 2A was selected as the embedded water molecule for this test. The remaining water molecules were represented by the same sets of point charges described in Section 3.2, and the M06-2X/MIDI! density matrix of the water molecule in each charge set was calculated. From this density matrix, one can calculate the dipole of the water molecule in two different ways: (a) quantum mechanically as the expectation value of the dipole moment operator or (b) classically from a set of point charges that has been determined through charge analysis of the water molecule's density matrix. Dipoles computed by methods a and b are called density dipoles and point-charge dipoles, respectively.

Method b begs the question of which specific charge model should be used to assign point charges to the embedded water molecule. To decide which charge model to use for this specific task, the dipoles of several nonembedded gas-phase molecules and ions according to nine charge models listed in Table 8. The ChElPG, ESP-Dipole, MK, and NBO charge analyses were carried out on the molecules' M06-2X/cc-pV(T+d)Z+ density matrices, the CM4M charges were determined from the M06-2X/MIDI! density matrices, and the CM1A, scaled CM1A, CM2, and CM3 charges were extracted from the AM1 wave functions. The geometries of these compounds had previously been optimized at the same level of theory at which the charge analyses were performed, except for the CM4M charges which had been optimized at M06-2X/cc-pV(T+d)Z+. The classical dipoles of these molecules calculated from these charge representations are shown in Table 8 and are compared to our best estimates of these dipoles. In the case of the neutral compounds, our best estimates are based on experimental values, but in the case of ions these have been determined with respect to each given ion's center of nuclear charge by finite-field calculations done with the CCSD(T)/aug-cc-pV(T+d)Z//M06-2X/cc-pV(T+d)Z+ method. The MUEs and RMSEs over all five compounds with respect to the best-estimate dipoles are also given in Table 8. Based on the MUEs, the CM4M charges appear to reproduce the best estimate dipoles better than the other methods, so the CM4M charges of the embedded water monomer were used to calculate the classical dipoles of method b.

The dipoles arising from methods a and b are given in Table 9, along with the individual point charges used to calculate the dipoles for method b. One should first notice, by comparing the point charge dipole of the unembedded water molecule (in the row labeled "None") to the point-charge dipoles of the embedded water molecule, that the electrostatic embedding does increase the dipole of the water molecule by about 0.3 D. Once embedded, however, the specific background charge set chosen does not affect the point-charge dipole by more than 0.05 D (or the density dipole by more than 0.09 D). The relative standard deviations over all sets of embedding charges for the water molecule's CM4M charges, density dipoles (method a), and point-charge dipoles (method b) are all under 1.3%. Thus, the extent to which a water molecule is polarized is affected by whether or not the water molecule is embedded in point charges, but the extent of polarization does not depend heavily on the specific set of embedding charges used.

5. Conclusions

The primary goals of this paper were (i) to test the overall accuracy of the EE-MB approximation for clusters involving both large and small monomers as well as a mix of ions and neutral molecules and (ii) to observe the dependence of the EE-MB approximation's accuracy on the background charges used as the electrostatic embedding.

Regarding the first goal, this study shows that the EE-MB approximation is capable of providing accurate binding energies for relatively complicated systems. For five sets of embedding charges used with three different basis sets on a test set of mixed clusters ranging in size from two to nine molecules, the errors in the binding energies from the EE-PA approximation relative to the binding energies of the full calculations at the same level of theory do not exceed 10%, and in many cases are closer to 5%. The EE-3B approximation does even better, with relative errors that do not exceed 0.7%.

Regarding the second goal, this study shows (in accord with results from previous studies on less complicated systems[13,16]) that electrostatic embedding does significantly improve the performance of the PA and 3B approximations, but that the specific set of point charges used for the electrostatic embedding does not strongly influence the accuracy of the EE-PA or EE-3B approximations. Two general categories of background charge sets were tested: geometry-dependent (GD) and geometry-independent (GI) charge sets. On the whole, GD and GI charges yield EE-MB results of almost equal accuracy; over three configurations of a mixed nine-molecule system, the EE-3B MUE over

Table 9. CM4M Charges (in *e*) and Dipoles (in debye) of Monomer Number 5 from Configuration 2A Embedded in the Given Sets of Background Charges

background charges	H1[a] (*e*)	O (*e*)	H2[b] (*e*)	density dipole[c] (D)	point-charge dipole[d] (D)
full[e]	0.321	−0.593	0.321	N/A[e]	1.87[e]
none[f]	0.297	−0.596	0.298	1.95	1.76
GI_ChElPG	0.371	−0.687	0.317	2.61	2.05
GI_ESP-dipole	0.371	−0.687	0.317	2.61	2.05
GI_MK	0.370	−0.687	0.317	2.61	2.05
GI_NBO	0.376	−0.690	0.314	2.63	2.06
GI_CM4M	0.366	−0.684	0.318	2.59	2.04
GI_CM1A	0.370	−0.686	0.315	2.60	2.05
GI_CM1A*1.14	0.380	−0.697	0.318	2.68	2.09
GI_CM2	0.377	−0.693	0.316	2.65	2.07
GI_CM3	0.375	−0.692	0.316	2.64	2.07
GI_SM5.42/CM2	0.377	−0.692	0.315	2.64	2.07
GI_SM8/CM4M	0.367	−0.684	0.317	2.59	2.04
GI_SMD/CM4M	0.368	−0.684	0.317	2.59	2.04
GD_Mulliken	0.378	−0.690	0.312	2.63	2.07
GD_CM1A	0.371	−0.684	0.313	2.59	2.05
GD_CM1A*1.14	0.380	−0.696	0.316	2.68	2.08
GD_CM2	0.377	−0.689	0.311	2.62	2.06
GD_CM3	0.377	−0.690	0.314	2.63	2.07
Average[g]	0.374	−0.689	0.315	2.62	2.06
standard deviation[g]	0.005	0.004	0.002	0.03	0.01
% standard deviation[g]	1.2	0.6	0.7	1.1	0.7

[a] The hydrogen atom (of monomer 5, see labels in Figure 2) that forms an H-bond with an oxygen atom of the nearby HSO_4^- ion. [b] The hydrogen atom (of monomer 5) that is not involved in an H-bond. [c] Dipole calculated from the M06 2X/MIDI! wave function of the embedded water molecule. One D ≡ 1 debye. [d] Dipole calculated from the point charges given in the columns labeled H1, O, and H2. One D ≡ 1 debye. [e] The CM4M charges assigned to the water molecule when the M06-2X/MIDI! calculation is performed on the entire 2A configuration. Because in this case the sum of the point charges on this fragment is not zero, the point charge dipole moment was calculated with respect to this fragment's center of nuclear charge. [f] "None" indicates that water monomer 5 was left in the geometry that it has in configuration 2A but that it was not embedded in point charges. [g] The values found in the rows labeled "full" and "none" were not included in the calculations of the averages, standard deviations, or % standard deviations.

the GI charge sets is 0.21 kcal/mol and over GD charge sets is 0.18 kcal/mol. Although the GD charge sets perform slightly better, they are also slightly more complicated to implement for energies and much more complicated to implement for gradients. Of the GI charge sets, those that were obtained from gas-phase monomer density matrices do not perform significantly differently than those that were obtained from liquid-phase monomer density matrices: over the EE-3B binding energies of three configurations of the nine-molecule system, the gas-phase monomer-derived charge sets yielded an MUE of 0.21 kcal/mol, and the liquid-phase monomer-derived charge sets yielded an MUE of 0.20 kcal/mol.

A third objective that arose during the course of this study was to investigate why the GD charge sets do not perform as much better than the GI charge sets as one might have expected. The conclusions reached from that portion of the study are these: (1) The polarization of the water molecules within a given nine-molecule cluster does not vary much from molecule to molecule, implying that the "rigid" point charges from a gas-phase monomer are adequate to represent that type of monomer regardless of where it is located within a cluster. (2) The polarization of a water molecule is affected by the presence of embedding charges, but the specific set of embedding charges used does not strongly affect the extent of the water molecule's polarization. The purpose of the background charges in an EE-MB calculation is to include higher-order effects in lower orders of the many-body expansion through the polarization of individual monomers and groups of monomers. Because the extent to which a water molecule is polarized is not greatly influenced by the choice of embedding charges, one can understand why the overall accuracy of the EE-MB approximation is not greatly influenced by the choice of embedding charges.

Table 10. Mean Unsigned Errors in kcal/mol over Three Configurations of an $(H_2SO_4)(HSO_4^-)(NH_4^+)(H_2O)_6$System[a]

	MB	EE-MB
one-body approximation	205.9	231.1
two-body approximation	9.8	3.0
three-body approximation	0.7	0.2

[a] M06-2X/cc-pV(T+d)Z+; see Table 3. The EE-MB results in the present table are averaged over the 12 geometry-independent charge models of Table 3.

As far as implications for future work, the most significant result of the present study is shown in Table 10, which shows results for the most complex systems in this article, namely the three 9-mers, each consisting of six water molecules, one ammonium ion, one bisulfate ion, and one sulfuric acid molecule. We see that the errors in the EE-MB approximation are very small, even for this complex cluster, and even with geometry-independent partial charges.

A concise summary of the major conclusions reached by this study is as follows: the EE-3B approximation as it was originally formulated (i.e., using geometry-independent background charges derived from equilibrium gas-phase monomer wave or density functions) can be trusted to provide accurate results for relatively complicated systems of widely varying sizes involving both ions and noncovalently interacting monomers.

Acknowledgment. The authors are grateful to the Minnesota Supercomputing Institute (www.msi.umn.edu) for computer time. This work was supported in part by the National Science Foundation grant no. CHE07-04974.

Supporting Information Available: Tables listing the Cartesian coordinates of the eight structures shown in Figure 1 are available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

(2) Shavitt, I. In *Methods of Electronic Structure Theory*; Schaefer, H. F. I., Ed.;Plenum: New York, 1977; pp 189−275.

(3) Bytautas, L.; Matsunaga, N.; Nagata, T.; Gordon, M. S.; Reudenberg, K. *J. Chem. Phys.* **2007**, *127*, 204301.

(4) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(5) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

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

(7) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 44903.

(8) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

(9) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193.

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

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

(12) Hirata, S. *J. Chem. Phys.* **2008**, *129*, 204104.

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

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

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

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

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

(18) Kulmala, M. *Science* **2003**, *302*, 1000.

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

(20) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(21) Baker, J. *Theor. Chim. Acta* **1985**, *68*, 221.

(22) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *PhysChemComm* **2002**, *5*, 117.

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

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

(25) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. J. Chem. Theory Comput., in press.

(26) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.

(27) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 872.

(28) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.

(29) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 10707.

(30) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291.

(31) Udier-Blagovicì, M.; Morales de Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322.

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

(33) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

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

(35) 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 Online Manual. http://www.gaussian.com/g_ur/g03mantop.htm (accessed Feb 13, 2009).

(36) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211.

(37) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2005**, *113*, 133.

(38) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2046.

(39) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(40) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.

(41) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. J. Phys. Chem. B, in press.

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

(43) Hariharan, P. C.; Pople, J. A. *Chem. Phys. Lett.* **1972**, *16*, 217.

(44) Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. *J. Chem. Phys.* **1998**, *109*, 1223.

(45) Zhao, Y.; Truhlar, D. G. Minnesota Gaussian Functional Module, version 3.0; University of Minnesota, Minneapolis, MN, 2007.

(46) Frisch, M. J., Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.;

Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03, revision D.01; Gaussian, Inc., Wallingford, CT, 2004.

(47) Olson, R. M.; Marenich, A. V.; Chamberlin, A. C.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Hawkins, G. D.; Winget, P. D.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J. Minnesota Gaussian Solvation Module, version 2008; University of Minnesota, Minneapolis, MN, 2008.

(48) Marenich, A. V.; Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. GESOL, version 2008; University of Minnesota, Minneapolis, MN, 2008.

(49) Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J.; Zhu, T.; Thompson, J. D.; Winget, P.; Lynch, B. J.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. AMSOL, version 7.1; University of Minnesota, Minneapolis, MN, 2004.

(50) Dahlke, E. E.; Truhlar, D. G. MBPAC, version 2007-2; University of Minnesota, Minneapolis, MN, 2007.

(51) Larson, L. J.; Largent, A.; Tao, F. *J. Phys. Chem. A* **1999**, *103*, 6786.

(52) Anderson, K. E.; Siepmann, J. I.; McMurry, P. H.; Vande-Vondele, J. *J. Am. Chem. Soc.* **2008**, *130*, 14144.

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

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

(55) Dyke, T. R.; Muenter, J. S. *J. Chem. Phys.* **1973**, *59*, 3125.

(56) Clough, S. A.; Beers, Y.; Klein, G. P.; Rotham, L. S. *J. Chem. Phys.* **1973**, *59*, 2254.

(57) Shostak, S. L.; Ebenstein, W. L.; Muenter, J. S. *J. Chem. Phys.* **1991**, *94*, 5875.

(58) Sedo, G.; Schultz, J.; Leopold, K. R. *J. Mol. Spectrosc.* **2007**, *251*, 4.

(59) Iwahori, J.; Ueda, Y.; Nakagawa, K. *J. Mol. Spectrosc.* **1986**, *117*, 1.

CT900095D

JCTC Journal of Chemical Theory and Computation

Symmetry-Adapted Perturbation Theory Applied to Endohedral Fullerene Complexes: A Stability Study of $H_2@C_{60}$ and $2H_2@C_{60}$

Tatiana Korona,*,† Andreas Hesselmann,*,‡ and Helena Dodziuk*,§

Faculty of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland, Institut für Physikalische und Theoretische Chemie, Universität Erlangen, Egerlandstrasse 3, 91058 Erlangen, Germany, and Institute of Physical Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44, 01-224 Warsaw, Poland

Received March 7, 2009

Abstract: Because of difficulties in a description of host−guest interactions, various theoretical methods predict different numbers of hydrogen molecules which can be inserted into the C_{60} cavity, ranging from one to more than 20. On the other hand, only one H_2 molecule inside the C_{60} fullerene has been detected experimentally. Moreover, a recently synthesized $H_2@C_{70}$ complex prevails in the mixture formed with $2H_2@C_{70}$. To get a deeper insight into the stability of the complexes created from C_{60} and hydrogen molecules, we carried out highly accurate calculations for complexes of one or two hydrogen molecules with fullerene applying symmetry-adapted perturbation theory (SAPT) and a large TZVPP basis set for selected points on the potential energy surfaces of $H_2@C_{60}$ and $2H_2@C_{60}$. The electron correlation in the host and guests has been treated by density functional theory. Our calculations yield the stability of the recently synthesized $H_2@C_{60}$ complex. In addition, for all tried positions of the H_2 dimer inside the C_{60} cage, the $2H_2@C_{60}$ complex has been characterized by a positive interaction energy corresponding to the instability of this species. Contrary to the conclusions of several theoretical studies, this finding, as well as model considerations and literature experimental data, indicates that only one hydrogen molecule can reside inside the C_{60} cage. The calculated energy components have been analyzed to identify the most important contributions to the interaction energy. Supermolecular interaction energies obtained with MP2, SCS-MP2, and DFT+Disp methods are also reported and compared to those of DFT-SAPT. The DFT-SAPT interaction energy has also been calculated for several points on the potential energy surface for a larger $2H_2@C_{70}$ complex, confirming, in agreement with recent experimental findings, that this species is stable. The DFT-SAPT approach has been used for the first time to obtain interaction energies for van der Waals endohedral complexes, demonstrating that the method is capable of handling such difficult cases.

Introduction

The possibility of filling a fullerene cage with atomic, ionic, or molecular guests was postulated soon after the serendipitous discovery of C_{60}[1] and development of the procedure for its purification.[2] Since then numerous proposals of possible applications of endohedral fullerene complexes[3] have appeared in different areas of science and technology, ranging from medicine[4] and environmental protection[5] to molecular optoelectronics[6] and renewable energetics.[7] Interestingly, almost none of them have been marketed yet, indicating the immense complexity of the task of filling and subsequent releasing guest molecules from the fullerene cavity. The utilization of fullerenes and carbon nanotubes as hydrogen storage devices has been recently one of the hot topics of research in view of their prospective practical applications.[8,9] However, the newest experimental evidence damped the expectations for storing a considerable amount of hydrogen inside these carbon forms.[10]

* Corresponding author e-mail: tatiana.korona@chem.uw.edu.pl; andreas.hesselmann@chemie.uni-erlangen.de; dodziuk@ichf.edu.pl.

† University of Warsaw.

‡ Universität Erlangen.

§ Polish Academy of Sciences.

10.1021/ct900108f CCC: $40.75
Published on Web 05/18/2009

Similarly to the case of nanotubes, a hydrogen molecule can be either chemi- or physisorbed on the fullerene surface. The chemisorption consists in a hydrogenation of the fullerene (i.e., the covalent CH bonds are formed). During the physisorption, a hydrogen molecule is attached to the C_{60} ball without the covalent bond formation. In the latter case, an endohedral van der Waals complex (denoted as $H_2@C_{60}$) is formed if H_2 is placed inside the fullerene, otherwise a more usual, but weaker exohedral complex is obtained. A synthesis of $H_2@C_{60}$ by Komatsu et al.[11] consisted in a chemical creation of a hole in the fullerene cage, insertion of a hydrogen molecule,[12] followed by a chemical closure of the cage.[11] This process has been called a "molecular surgery" of fullerenes.[12−14] Note that a parting with an idea of a "brute force" insertion of hydrogen inside the fullerene under extreme conditions (high temperature and pressure) in favor of a milder chemical opening of the cage was suggested some time ago by Patchkovskii and Thiel for $He@C_{60}$[15] and by Dodziuk et al.[16] Very recently, Murata et al.[17,18] used a similar approach to obtain the $2H_2@C_{70}$ complex. It is noteworthy that the experimentally observed relative population of $H_2@C_{70}$ and $2H_2@C_{70}$ (97:3)[18] is very close to the values which can be deduced from a difference of 1.8 kcal/mol in the steric energy values determined by molecular mechanics.[19]

Two main issues have been addressed by theoreticians when studying endohedral complexes of fullerenes with hydrogen molecules: a height of the barrier hindering the guest from entering the fullerene cavity and an estimation of the number of hydrogen molecules which can be hosted by the fullerene cage. Both problems can be first analyzed on the basis of simple mechanistic considerations. By taking the length of the H−H bond of about 0.7 Å, the van der Waals radii of the H and C atoms of ca. 1.2 and 1.7 Å, respectively, and assuming the diameter of fullerene (treated approximately as a sphere) as 7.1 Å,[20] it is easy to estimate that there is no room left for another H_2 molecule inside the cage. Such a simple reasoning shows that only one hydrogen molecule can be hosted by C_{60}. Analogously, a comparison of the radii of the H and C atoms allows one to draw a conclusion that a hydrogen atom or molecule passing through the five- or six-membered ring should exhibit a strong repulsion because of an overlap of the electron clouds of the ring and the H atoms.

The problem of a barrier estimation for a guest entering the C_{60} cage has been studied in several articles,[21−23] yielding barriers of 3 eV[21] or 2 eV[23] for a hydrogen atom and 20 eV for H_2,[22] when passing through the six-membered ring. Subsequent molecular dynamics simulations have shown a very low probability for the process of catching a hydrogen atom inside the cage, while the same process in the case of the hydrogen molecule has been not observed in silico at all.[21] These calculations are in line with a lack of success in obtaining $H_2@C_{60}$ by a direct hydrogen insertion into an intact C_{60} fullerene (i.e., without opening the cage).[11,18]

The stability of complexes of hydrogen molecule(s) buried inside the C_{60} cage has been studied by various approaches, from molecular mechanics (MM),[24] through semiempirical and density-functional theory (DFT), to ab initio quantum chemical (QM) methods. At the beginning, it should be stressed that a usage of the semiempirical approach for nonbonding interactions is rather counterproductive, since this method has been developed to provide approximate energies of chemically *bonded* systems, and for this very reason the calculations of *nonbonded* complexes performed with this class of methods cannot be reliable. One can also add that semiempirical methods are known to have difficulties even with a satisfactory description of hydrogen bonds,[25] which are by orders of magnitude stronger than interactions between a hydrogen molecule and a π-electron system. Therefore, the reports of Turker and Erkoc,[26] who found a stabilization of 24 hydrogen molecules inside C_{60} on the basis of semiempirical AM1 calculations, or those of Ren et al.,[27] who used the PM3 method combined with DFT and inserted 25 H_2 into the cage, are not reliable. A detailed criticism of the results of Turker and Erkoc can be found in refs 28 and 29. It should be noted that a recently developed OM*x* class of semiempirical methods compares somewhat better with QM calculations, but still without an empirical dispersion correction they predict no interaction for complexes containing π-electron systems.[30]

The accuracy of quantum chemical calculations for endohedral fullerene complexes is usually quite limited. The obvious reason for this state of affairs is the size of the system,[31] which precludes the use of the high-level ab initio electron-correlated theories, such as, for example, coupled cluster[32] (see, however, ref 33) and large orbital basis sets. The limitations in the choice of ab initio electron-correlated theories are especially serious in the case of nonpolar or slightly polar guest molecules, since in these cases the host−guest interactions are mainly dispersive and exchange ones while it is known that the dispersion energy (and the corresponding exchange-dispersion energy) is not accounted for by the Hartree−Fock method.[34] For instance, early self-consistent field (SCF) calculations of Cioslowski for $H_2@C_{60}$ yielded incorrectly the instability of this complex.[35]

Another popular method of accounting for the electron correlation is the DFT approach. The commonly employed DFT in its local density approximation or generalized gradient approximation variant is not an ab initio method, and the accuracy of reproducing the electron energy is strongly dependent on the quality of the functionals used. It is well-known that the DFT approach with commonly used functionals often strongly underestimates[36] the stabilizing dispersion interaction (and a smaller destabilizing exchange-dispersion contribution) and that a neglect of this part of the interaction energy can lead to qualitatively incorrect results (e.g., destabilization instead of the stabilization effect). It should be also stressed that the existence of a multitude of DFT functionals makes it very difficult for a nonspecialist to select the best one for his or her particular purpose, although recent comprehensive studies on the quality of various functionals for several classes of nonbonding interactions provide some general guidelines.[37] This situation results, unfortunately, in a broad misuse of functionals, leading to many erroneous conclusions. However, at present the reason why most functionals are incapable to describe dispersion interactions is well-known[38] and several solutions

to this problem exist, the most sophisticated of which are possibly the ones which derive nonlocal orbital functionals from many-body perturbation theory,[39] coupled cluster theory,[40] or the fluctuation−dissipation theorem DFT.[41] Another practical solution to the dispersion energy problem with DFT is to use empirical damped multipole expansions of the dispersion energy as a correction term to the DFT energy (see, for example, refs 42−44). In these methods, the dispersion coefficients are often calculated as combinations of atomic dispersion coefficients and thus the dispersion energy can be estimated very efficiently. It has to be added, though, that the dispersion energy obtained in this way should not be confused with the dispersion energy from an intermolecular perturbation theory and has in fact no physical meaning. To our best knowledge, the most popular method from this class (i.e., the DFT+Disp method of Grimme[42,43]) has not been used to study the stability of endohedral complexes of fullerenes yet. Another modification of the standard DFT method, known as DFT with tight binding (DFTB+),[44] correctly yields only one hydrogen molecule stabilized inside the C_{60} fullerene;[45] however, the latter authors claimed the stability of the highly strained endohedral complexes with up to 20 H_2, which, according to model considerations and available experimental results, cannot be obtained.

The existence of some new promising DFT functionals, which are especially designed to imitate the nonbonding interactions, should be also mentioned here. One of these functionals (MPWB1K,[46] thorougly tested by Zhao and Truhlar[37]) has been recently used by Slanina et al.[47] to estimate the stabilization energies of H_2, Ne, and N_2 inside the C_{60} cavity. In agreement with the experimental evidence,[11,48] MPWB1K predicts that these complexes are stable. The values of the MPWB1K stabilization energies are similar to values obtained from the second-order Møller−Plesset (MP2) and spin-component-scaled MP2 (SCS-MP2)[49] calculations. However, as the authors of ref 47 notice, the MPWB1K functional does not provide a correct description of stacking interactions (like those between two benzene molecules), and therefore it cannot be excluded that sensible values of the interaction energies for three endohedral complexes studied there are just a result of an accidental error cancellation. Additionally, it is known that for complexes involving aromatic molecules the MP2 method often gives too large values of the attractive interaction energies,[50] and therefore a good agreement with MP2 cannot be viewed as an ultimate proof of the usefulness of the MPWB1K functional for such cases.

Unfortunately, DFT with standard functionals is still utilized to calculate energies of endohedral fullerene complexes without taking into account a missing dispersion component of the interaction energy. Among several such works dealing with the stabilization effect for hydrogen molecules in the C_{60} fullerene, one can list, for example, articles of Yang,[51] Pupysheva et al.,[9] and Lee and McKee.[52] These authors claim to find stable[51,52] or metastable[9] structures involving numerous H_2 molecules (plus eventually partly chemisorbed species) inside C_{60} using standard DFT functionals, although in the Yang and Lee and McKee articles starting from the second added hydrogen molecule the energy of the complex is higher than the sum of the energies of isolated molecules. In the work of Pupysheva et al., two H_2 molecules are stabilized inside C_{60}, and for the number of hydrogen molecules in the fullerene cage greater or equal to 15 a partial chemisorption has been obtained yielding unphysically long CH bond of even 1.20 Å. Yang inserted up to 29 hydrogen molecules into the C_{60} cage and claimed that only for 29 guests the cage will be broken. He also modeled the hydrogen entrance into the cage, stating that 19 H_2 molecules can pass through a small opening involving nine bonds. This result contradicts the experimental studies[53] on the orifice size enabling the entrance of one hydrogen molecule inside C_{60}. The paper by Yang[51] has been criticized by Dolgonos,[54] who pointed to the unreliability of the DFT calculations in this case and to very short distances between the seemingly "nonbonded" hydrogen atoms. The Yang reply to the comment of Dolgonos has been unsubstantial.[55] Lee and McKee studied the reactivity of up to six H_2 molecules inside C_{60} using DFT and MP2 methods with unreliably small basis sets. Also, HH distances of 1.6 Å reported by Lee and McKee[52] are certainly too small and should lead to a considerable repulsive destabilization of the systems under consideration. The analyses by the latter authors and Pupysheva et al. of the pressure inside the fullerene cage filled with numerous hydrogen molecules seem immaterial since, as discussed earlier, these complexes cannot be realized. It should be stressed that if an endohedral complex with two or even more endohedral H_2 molecules had been formed, then, despite a high strain, it would not decompose unless the strain of the complex distributed over the whole cage would be sufficiently large to break it. However, no process that could provide complexes with more than one guest inside C_{60} seems feasible. On the other hand, recent claims[9,52] that endohedral fullerene complexes with hydrogen molecules can be of use for hydrogen storage seem unfounded, since the release of guest hydrogen molecules should lead to an irreversible cage destruction. A recent idea to store hydrogen in chemically opened fullerene cages[56] could be a route to overcome this obstacle.

In this work, the endohedral C_{60} complexes involving one or two hydrogen molecules will be investigated using a computationally efficient variant of intermolecular symmetry-adapted perturbation theory (SAPT),[57] which allows one to reliably estimate the interaction energies in the $H_2@C_{60}$ and $2H_2@C_{60}$ species. A simultaneous study of these two complexes allows us to investigate a delicate balance between the dispersion and repulsion energies, which dominate in the intermolecular interactions for these two species, thus demonstrating the applicability of the latter method for such complicated cases.

Methods

Let us consider the interaction of two or three closed-shell molecules (denoted A, B, C). In general, the interaction energy of *m* molecules A, B, C,... is defined as a difference,

$$E_{\mathrm{int}}(\mathrm{ABC...}) = E_{\mathrm{ABC...}} - (E_{\mathrm{A}} + E_{\mathrm{B}} + E_{\mathrm{C}} + ...) \quad (1)$$

where $E_{\mathrm{A,B,C...}}$ is the energy of the complex ABC... and E_{X} is the energy of the molecule X (X = A, B, C, ...). The

interaction energy of the three molecules A, B, and C can be separated into the additive and nonadditive parts:

$$E_{\mathrm{int}}(\mathrm{ABC}) = E_{\mathrm{int}}[2,3] + E_{\mathrm{int}}[3,3] \quad (2)$$

where $[n,m]$ denotes the n-body contribution for the complex of m molecules. The additive part $E_{\mathrm{int}}[2,3]$ is thus defined as a sum of interaction energies of all pairs:

$$E_{\mathrm{int}}[2,3] = E_{\mathrm{int}}(\mathrm{AB}) + E_{\mathrm{int}}(\mathrm{BC}) + E_{\mathrm{int}}(\mathrm{CA}) \quad (3)$$

and the nonadditive part $E_{\mathrm{int}}[3,3]$ accounts for a modification of the interaction caused by the third partner. Note that in eqs 1−3 the intramolecular geometry parameters of A, B, and C have not been changed when calculating energies of complexes (i.e., no geometry relaxation is taken into account).

Equations 1−3 directly define the so-called supermolecular approach (sometimes called supramolecular one) for the calculation of interaction energies. In the supermolecular method, one calculates energies of all molecules and complexes (A, B, AB, etc.) by a given method and just makes the appropriate subtractions, according to eqs 1−3. Although appealing at first look, this approach has several disadvantages (see, for example, ref 58 for a detailed discussion). However, if a suitable theory is selected for the calculation of the electron energies and if the counterpoise correction of Boys and Bernardi[59] is used, the supermolecular approach can produce reliable potential energy surfaces (PES) for the van der Waals complexes.

It should be noted parenthetically that endohedral species such as $H_2@C_{60}$ are untypical examples of the van der Waals complexes since they cannot be separated into their constituent parts without the cage breaking. However, from the theoretical point of view there is no difference in a treatment of the endo- and exohedral van der Waals species.

SAPT Treatment of the Interaction Energy of Two Molecules. Another well-established approach for the calculation of the interaction energy for two closed-shell molecules is symmetry-adapted perturbation theory.[57,60] In SAPT, the interaction energy is obtained directly as a sum of well-defined physical contributions and *not* as a difference between two similar numbers (see eq 1). Up to the second order in terms of the intermolecular interaction operator $V = H_{\mathrm{AB}} - H_{\mathrm{A}} - H_{\mathrm{B}}$ (where H_{X} is the electron Hamiltonian of a molecule or a complex X,X = AB, A, B), these contributions comprise: the first-order electrostatics ($E^{(1)}_{\mathrm{elst}}$), second-order induction ($E^{(2)}_{\mathrm{ind}}$) and dispersion ($E^{(2)}_{\mathrm{disp}}$) energies, and their exchange counterparts: first-order exchange ($E^{(1)}_{\mathrm{exch}}$), second-order exchange-induction ($E^{(2)}_{\mathrm{exch\text{-}ind}}$) and exchange-dispersion ($E^{(2)}_{\mathrm{exch\text{-}disp}}$), accounting for the electron tunneling between the interacting constituent molecules. The SAPT method up to the second order in V gives the main part of the interaction energy. As an estimation of the higher-order induction and exchange-induction energies, the Hartree−Fock "delta" correction term δE_{HF} is usually utilized.[34,61] Summarizing, the interaction energy in SAPT is calculated as:

$$E^{\mathrm{SAPT}}_{\mathrm{int}} = E^{(1)}_{\mathrm{elst}} + E^{(2)}_{\mathrm{ind}} + E^{(2)}_{\mathrm{disp}} + E^{(1)}_{\mathrm{exch}} + E^{(2)}_{\mathrm{exch-ind}} + E^{(2)}_{\mathrm{exch-disp}} + \delta E_{\mathrm{HF}} \quad (4)$$

To calculate the energy contributions listed above, the exact wave functions of molecules A and B should be known in principle. Since usually these solutions are not available, one has to resort to some approximate methods. The simplest solution is the utilization of the Hartree−Fock (HF) determinants, in which case the so-called SAPT(HF) method is obtained. In this method, the effect of the electron correlation inside the A and B molecules is completely neglected. Thus far, three methods have been developed which enable to include the effect of the electron correlation inside the interacting molecules: (i) historically the first and the most popular SAPT(MP) approach,[62,63] where the molecules A and B are treated by Møller−Plesset (MP) theory, (ii) SAPT(CC) approach,[64] developed by Korona and Jeziorski, where these molecules are described at the coupled cluster level (see also early works[63,65]), and (iii) the SAPT method with intramolecular electron correlation described by DFT. Only the latter method can treat molecules of the fullerene size, and therefore it will be described below in more detail.

A possibility of using DFT to account for the intramolecular correlation in SAPT was first pointed out in ref 66. The formalism of the DFT-SAPT method has been developed independently in two groups: Hesselmann and Jansen[67,68] and Misquitta et al.[69] The implementation of DFT-SAPT, followed by a recent inclusion of the density-fitting (DF) formalism[70] for the calculation of two-electron repulsion integrals, allows one to extend treatable sizes of molecules by an order of magnitude. In particular, a DFT-SAPT calculation for a molecule of the C_{60} size has become feasible. The idea of DFT-SAPT consists in using the Kohn−Sham (KS) and the coupled-perturbed KS (CKS) orbitals instead of the HF and coupled-perturbed HF orbitals in SAPT(HF). In this way, the electron correlation of molecules A and B, present in DFT orbitals, is taken into account in SAPT at cost of the SAPT(HF) method. It should be stressed that DFT-SAPT is a different method from the supermolecular DFT and that the individual interaction energy terms in DFT-SAPT cannot be obtained from an energy decomposition of the supermolecular DFT energy. In particular, DFT-SAPT accounts correctly for the dispersion effect, since the dispersion and exchange-dispersion energies are calculated as the corresponding SAPT corrections. The accuracy of the DFT-SAPT method has been recently confirmed by a comparison with benchmark SAPT(CC) calculations[64] and with the supermolecular CCSD(T) approach (see, for example, ref 71).

Interaction Energies of Three Molecules. The SAPT method has been extended for the interaction of three molecules in ref 72. In this approach, apart from the calculation of the usual SAPT interaction energies for three pairs of the complexes (AB, BC, and CA), one has to obtain the nonadditive contributions to the interaction energy. However, the program which calculates these corrections is strongly limited to small molecules. Fortunately enough, it can be demonstrated that an approximate sum of some of these corrections is incorporated, along with some higher-order corrections, in the nonadditive part of the supermolecular interaction energies, calculated at various levels of the supermolecular approach. Recently, using this feature,

Podeszwa and Szalewicz[73] developed two hybrid schemes for calculating these contributions. Both schemes divide the nonadditive interaction energy into two parts: one calculated by the supermolecular approach and another part calculated by perturbation theory. For this study, we selected the scheme denoted in ref 73 as MP2+SDFT. In the MP2+SDFT approach, the nonadditive part of the interaction energy is calculated as a sum of the MP2 supermolecular nonadditive interaction energy $E_{\mathrm{int}}^{\mathrm{MP2}}[3,3]$ and the perturbational three-body dispersion energy $E_{\mathrm{disp}}^{(3)}(\mathrm{CKS})[3,3]$, calculated from the CKS propagators of constituent molecules

$$E_{\mathrm{int}}[3,3] = E_{\mathrm{int}}^{\mathrm{MP2}}[3,3] + E_{\mathrm{disp}}^{(3)}(\mathrm{CKS})[3,3] \quad (5)$$

It was stated in ref 73 that the $E_{\mathrm{int}}^{\mathrm{MP2}}[3,3]$ term provides an estimation for the following nonadditive contributions: first-order exchange, second- and higher-order induction and exchange-induction, and a third-order mixed induction-dispersion terms. The third-order dispersion correction $E_{\mathrm{disp}}^{(3)}$ is absent in the supermolecular MP2 method, and it should be therefore calculated separately. It should be stressed that at least third-order Møller−Plesset theory (MP3) is required to account for $E_{\mathrm{disp}}^{(3)}$, which for the nonpolar species is a dominant nonadditive long-range effect. Summarizing, the total interaction energy in the hybrid scheme is obtained as a sum of the following contributions:

$$E_{\mathrm{int}}^{\mathrm{hybrid}} = E_{\mathrm{int}}^{\mathrm{SAPT}}(\mathrm{AB}) + E_{\mathrm{int}}^{\mathrm{SAPT}}(\mathrm{BC}) + E_{\mathrm{int}}^{\mathrm{SAPT}}(\mathrm{CA}) + E_{\mathrm{int}}^{\mathrm{MP2}}[3,3] + E_{\mathrm{disp}}^{(3)}(\mathrm{CKS})[3,3] \quad (6)$$

Computational Details

All calculations were performed with the development version of the MOLPRO suite of programs.[74] In addition to the DFT-SAPT calculations, supermolecular calculations were performed with the MP2, SCS-MP2,[49] and dispersion-corrected DFT functional using the damped multipole expansion scheme developed by Grimme[43] to assess the quality of these methods in comparison to DFT-SAPT. The Boys−Bernardi counterpoise correction was used for all supermolecular calculations.[59]

DFT Calculations for a Fullerene and a Hydrogen Molecule. The C_{60} and H_2 molecules in DFT-SAPT were treated with the PBE functional[75] using an additional asymptotic correction of the exchange-correlation (xc) potential, as proposed by Grüning et al.[76] The utilization of this correction is crucial in this method, since otherwise the asymptotic density is in general too diffuse, leading to a poor description of magnitudes of intermolecular interactions.[67,69] This asymptotic correction is currently performed using a scheme which connects the respective xc potential in the bulk region with an asymptotic xc potential (having a Coulombic $-1/r$ behavior) by shifting the bulk potential by the so-called derivative discontinuity (i.e., the difference between (negative) ionization potential and HOMO energy of the underlying xc functional). For the case of the C_{60} molecule, the value of this correction was set to 0.0641 hartree and for the hydrogen molecule to 0.185 hartree. These values were obtained from the experimental vertical ionization potentials of C_{60} (0.279 hartree)[77] and H_2 (0.566 hartree)[78] and the corresponding HOMO energies of both systems using the PBE xc functional in the TZVPP basis set (−0.215 and −0.381 hartree, respectively). The latter functional was also used in the DFT+Disp method.[43]

A total nonadditive contribution to the interaction energy was calculated by the MP2+SDFT method. Additive (i.e., two-body) contributions were calculated by DFT-SAPT. Because of the absence of the basis-set superposition error[59] in the perturbational approach, the $H_2\cdots H_2$ and $H_2@C_{60}$ interaction energies in the $2H_2@C_{60}$ complex can be calculated without using the basis on the ghost molecule. In this way, we can utilize the results from the $H_2@C_{60}$ calculations. The additive contributions of the third order were neglected in the present study, unless they are present in the δE_{HF} term.

The core electrons (1s) for carbon atoms were frozen in all correlated calculations.

Choice of the Basis Set and Complex Geometries. The selection of a proper orbital basis set is crucial to obtain reasonable results. Because of the size of the system, we had to find a balance between the accuracy and the computational cost of the method. After some testing, we found that the TZVPP basis set[79,80] is the smallest reliable basis for our purposes. The corresponding cc-pVTZ/JKFIT[81] DF auxiliary basis set was used for the calculation of Coulombic and exchange integrals in SCF and the first-order interaction energy contributions while all doubly external integrals and all xc-type integrals occurring in the second-order DFT-SAPT were computed using the TZVPP/MP2FIT[82] fitting basis set. With these basis sets, the calculations for a single DFT-SAPT point (without the δE_{HF} correction) take about 5.5 days on Opteron/2 GHz and 2.5 days on Woodcrest/2.4 GHz computers.

The CC bond lengths of 1.458 and 1.401 Å were assumed[83] for the bonds in a pentagon ring and those between pentagon rings which, due to the I_h symmetry, fully determine the C_{60} geometry. As recommended in ref 84, we use the value of the vibrationally averaged $R_{\mathrm{H-H}}$ of 0.7668 Å. In view of the large size of the complexes under study, their full PES values could not be calculated. Instead, only few potentially interesting geometries of these two species were analyzed. For the $H_2@C_{60}$ complex, these geometries comprise three orientations relative to a selected pentagon ring of the fullerene (with the geometrical center of the hydrogen molecule lying on the fivefold symmetry axis of this pentagon), and two orientations related to a selected hexagon ring of the fullerene (with the geometrical center of the hydrogen molecule lying on the threefold symmetry axis of this hexagon). These orientations will be denoted as:

TP, a hydrogen molecule perpendicular to a selected pentagon ring;

PP, a hydrogen molecule parallel to a selected pentagon ring, H_2 lies in one of five symmetry planes of this pentagon;

SP, a hydrogen molecule forming the angle 45° to a selected pentagon ring; as in the case of the PP mutual orientation, H_2 lies in one of five symmetry planes of this pentagon;

TH, a hydrogen molecule perpendicular to a selected hexagon ring;

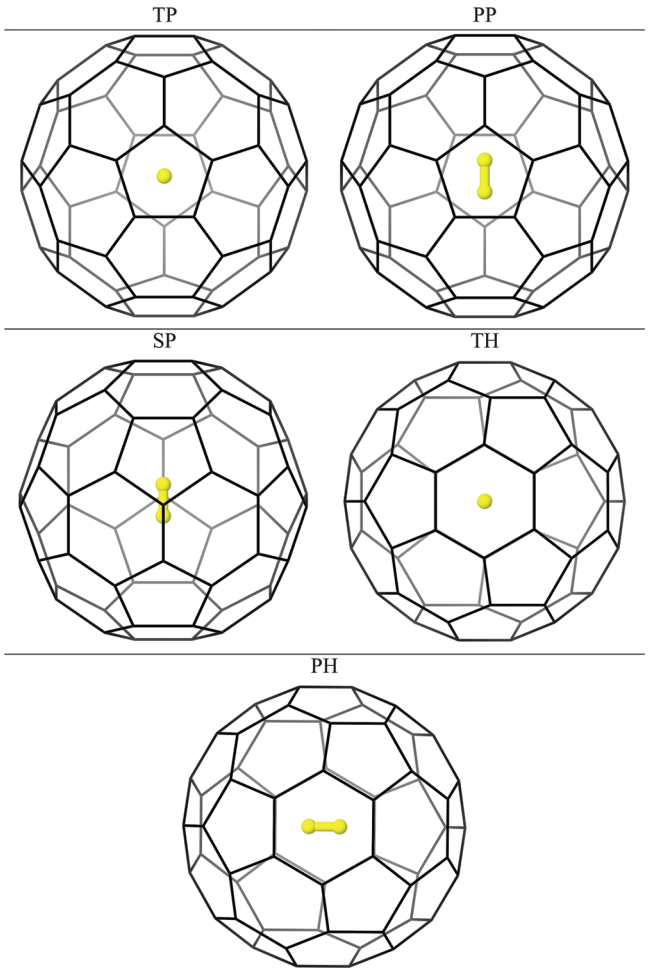


Figure 1. Studied orientations of one hydrogen molecule inside the fullerene molecule.

PH, a hydrogen molecule parallel to a selected hexagon ring, H_2 lies in one of three symmetry planes of this hexagon.

These orientations are also depicted in Figure 1. For each of these orientations, several distances r from the center of mass of the hydrogen molecule to the center of the fullerene were used. In two cases also distances $r > 3.5$ Å were taken into account, which correspond to the C_{60} complex with a hydrogen molecule outside the cage.

To select potentially interesting structures of $2H_2@C_{60}$, we first analyzed the CCSD(T) potential energy surface of the H_2 dimer, published recently by Hinde.[85] The global minimum for this system (−0.467 kJ/mol) occurs for a perpendicular (T) structure ($\theta_1 = 90°$, $\theta_2 = 0°$, coordinates defined in ref 85) at a distance of 3.36 Å between the geometrical centers of hydrogen molecules. To select revealing guest positions in the $2H_2@C_{60}$ complex, it is also important to know at which point the interaction energy of the $H_2\cdots H_2$ dimer is equal to zero. For the case of the T-structure, this happens at 2.92 Å. The minimum is only slightly shallower (−0.436 kJ/mol) for the skew (S) structure ($\theta_1 = 45°$, $\theta_2 = 45°$, $\phi = 0°$) with the zero point at 2.95 Å. We also found that it will be of interest to check two "crossed" orientations: X1 ($\theta_1 = 90°$, $\theta_2 = 90°$, $\phi = 72°$) and X2 ($\theta_1 = 90°$, $\theta_2 = 90°$, $\phi = 60°$). The selected structures of the H_2 dimer were inserted into the fullerene molecule, so that (i) both hydrogen molecules are equidistant from the center of C_{60}, (ii) they are placed at the orientations previously used for $H_2@C_{60}$, and (iii) their geometrical centers lie on the same symmetry axis of fullerene. In this way, the following structures with two guest molecules were selected (the unspecified details of geometries are the same as for the $H_2@C_{60}$ case):

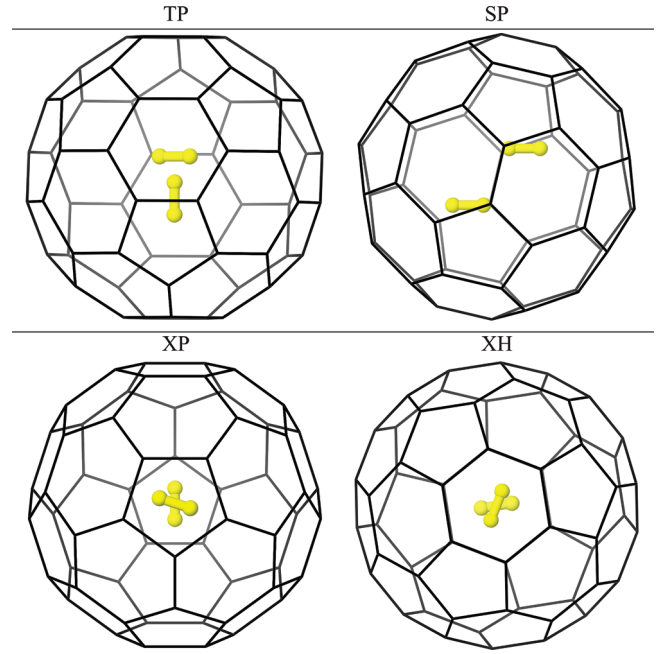


Figure 2. Studied orientations of two hydrogen molecules inside the fullerene molecule.

TP, a first hydrogen molecule parallel to a selected pentagon ring, the second one perpendicular to it (both H_2 forming the T-structure);

SP, both hydrogen molecules forming the 45° angle with a selected pentagon ring (both H_2 forming the S-structure);

XP, both hydrogen molecules parallel to a pentagon ring form the X1 structure;

XH, both hydrogen molecules parallel to a hexagon ring form the X2 structure.

These orientations are depicted in Figure 2.

Results and Discussion

C_{60} with One Hydrogen Molecule. The DFT-SAPT interaction energies for the complex of C_{60} with one H_2 molecule as a function of a distance from the cage center are presented in Table 1. The examination of this table reveals that there is a small stabilization effect for the endohedral complex of one hydrogen molecule with C_{60}. This effect is not large, since the minimum depth is equal to −19.35 kJ/mol, and already at $r = 1.0$ or 1.1 Å (depending on orientation), the interaction energy becomes positive. From the five orientations studied in this work, TH gives the lowest interaction energy, corresponding to the largest stabilization, although the differences between various orientations are very small, especially in the center of the fullerene cage. It is interesting to note that there is a shallow secondary minimum (or a saddle point) for the PH orientation, shifted by about 0.1 Å from the center. Another minimum region for this species occurs for the exohedral

Table 1. DFT-SAPT Interaction Energy for Selected Orientations of the $H_2@C_{60}$ Complex[a]

r/orientation	PP	TP	SP	TH	PH
0.00	−19.26	−19.26	−19.30	−19.35	−18.59
0.10	−19.25			−19.12	−19.14
0.20	−18.89				−18.68
0.25	−18.68	−18.79	−18.62	−18.87	−18.76
0.30	−18.52				−18.38
0.50	−17.25	−16.67	−16.78	−16.56	−17.08
0.75	−12.29	−10.78	−11.46	−11.23	−12.21
0.80	−9.88	−9.07			
0.90	−7.09	−4.41			−6.91
1.00	−1.74	1.71	−0.25	0.64	−1.99
1.10	5.44	9.90		8.30	4.97
1.25	19.67	27.19	23.12	24.22	19.42
1.50	62.61	76.00	68.37	68.18	59.88
1.75	142.55	163.39			
2.00	283.67	310.91			
5.00		173.49			
6.00	1.25	4.38			
6.50	−3.07	−2.24			
7.00	−1.64	−2.17			
8.00	−0.60	−0.82			

[a] Energy values in kilojoules per mole (1 millihartree = 2.6255 kJ/mol); distances in angstroms. Note that for distances $r > 3.5$ Å the complex is *exohedral*.

complex at $r \approx 6.5$ Å, but in this case the stabilization energy is too small to enable the complex stability at room temperature.

A detailed division of the SAPT interaction energy into components and interaction energies obtained by the supermolecular MP2, SCS-MP2, and DFT+Disp[43] approaches are presented in Table 2 for the PP orientation (a parallel orientation was selected for a more detailed analysis since it turns out that this orientation is preferred for the $2H_2@C_{60}$ case). The energy components for TH, TP, SP, and PH orientations are very similar to the presented ones and are given in the Supporting Information. A distance dependence of the SAPT corrections for the PP orientation is depicted in Figure 3.

Let us first focus on the supermolecular interaction energies presented in Table 2. An inspection of these data indicates that the MP2 method overestimates the complex binding, which is the common behavior of MP2 for the interaction involving aromatic rings,[50] while SCS-MP2 is in a much better agreement with DFT-SAPT. Next, let us look at the results of the DFT+Disp method of Grimme. It can be observed that DFT+Disp performs well at the center, but for larger distances $E_{\mathrm{int}}^{\mathrm{DFT+Disp}}$ increases less steeply than $E_{\mathrm{int}}^{\mathrm{SAPT}}$. The reason for such a behavior of the DFT+Disp method can be ascribed to the "dispersion" contribution of the latter method which rapidly decreases as the H_2 molecule approaches the cage wall. As a result, too much space is available for the hydrogen molecule according to the Grimme method. In the strong repulsive region of $r > 1.5$ Å, the correspondence of DFT+Disp with DFT-SAPT improves because of the switching on the damping function in the DFT+Disp method and a decrease of the dispersion contribution as compared to the other contributions in the supermolecular PBE interaction energy. This behavior is also found for other orientations studied and leads to the conclusion that the DFT+Disp method may not be accurate enough to study PES of endohedral hydrogen molecules in the C_{60} cage. However, we observed that the agreement of DFT+Disp with our DFT-SAPT reference data can be considerably improved by a modification of the damping parameter α from 20.0 to 9.2 and the prefactor $s6$ from −0.75 to −0.63 of the underlying original Grimme model.[43] While an application of this path may certainly not be advisible in general, it could provide a possible option to investigate the potential energy surface using a quantum chemistry method less expensive than MP2 or DFT-SAPT.

Let us analyze the behavior of the components of the SAPT interaction energy. The stabilization effect in the center of C_{60} comes mainly from the dispersion energy, while the first-order exchange energy gives the most important repulsive contribution. This trend continues as we approach the cage wall: both corrections grow in absolute values, but the dispersion effect increases slower, and finally the first-order exchange energy prevails leading to the repulsive character of the interaction. The induction energy is almost as important as the dispersion energy, but it is strongly damped by its exchange counterpart (this is a common effect for the short-range induction contribution; see, for example, ref 86). Nonetheless, for $r > 1.5$ Å the effective $E_{\mathrm{ind}}^{(2)} + E_{\mathrm{exch\text{-}ind}}^{(2)}$ contribution becomes more important than the dispersion energy.

The above analysis shows that great care should be exercised when modeling PES for endohedral fullerene complexes with a simple repulsion+dispersion model (see, for example, ref 87), since neglected short-range terms may become as large as the included ones, when approaching the cage wall.

An examination of Figure 3 reveals that for the PP orientation there is a shallow well in the attractive region and a steep repulsive potential wall for larger r, where the guest approaches the host cage. A similar pattern is found for other orientations. Table 1 shows that the center of the well is practically isotropic and large enough to allow for an almost free rotation of the H_2 guest. Anisotropy becomes more pronounced for larger distances (i.e., closer to the cage wall). A comparison of the data from the Supporting Information allows us to conclude that, as expected, the dispersion and exchange-dispersion energies are the most isotropic SAPT terms, while the first-order exchange, induction, and exchange-induction energies exhibit the largest anisotropy. However, even at $r = 1.5$ Å (highly repulsive region) this anisotropy does not exceed a few percent (e.g., first-order exchange corrections for the PH and TP orientations differ by 18% for this distance).

Let us analyze how the just presented results can be used to select the most interesting geometries describing the $2H_2@C_{60}$ complex. In view of the data from Table 1, shifting of a hydrogen molecule from the center by more than 1.0 Å will cause a strong repulsion from carbon atoms. This means that two hydrogen molecules in the fullerene cage can be separated by at most 2 Å, otherwise a strong repulsion from the cage wall will result. However, the PES for two hydrogen molecules is highly repulsive for such a small distance.[85] On the other hand, the PES for the H_2 dimer passes through zero at about 3 Å. If two hydrogen molecules are placed on

Table 2. Components of the DFT-SAPT Interaction Energy for the PP Orientation of the $H_2@C_{60}$ Complex[a]

r	$E^{(1)}_{elst}$	$E^{(1)}_{exch}$	$E^{(2)}_{ind}$	$E^{(2)}_{exch\text{-}end}$	$E^{(2)}_{disp}$	$E^{(2)}_{exch\text{-}disp}$	δE_{HF}	E^{SAPT}_{int}	E^{MP2}_{int}	$E^{SCS-MP2}_{int}$	$E^{DFT+Disp}_{int}$
0.00	−7.20	21.16	−5.02	4.50	−36.09	4.13	−0.74	−19.26	−30.69	−21.58	−21.27
0.10	−7.34	21.54	−5.16	4.63	−36.38	4.21	−0.76	−19.25	−30.65	−21.52	
0.20	−7.77	22.70	−5.54	4.99	−36.81	4.35	−0.80	−18.89	−30.54	−21.31	
0.25	−8.10	23.58	−5.82	5.24	−37.20	4.46	−0.85	−18.68	−30.44	−21.14	−21.11
0.30	−8.49	24.65	−6.16	5.55	−37.81	4.62	−0.89	−18.52	−30.31	−20.93	
0.50	−11.12	31.71	−8.66	7.79	−41.34	5.54	−1.17	−17.25	−29.29	−19.39	−20.60
0.75	−17.58	48.81	−14.66	13.15	−47.59	7.39	−1.81	−12.29	−25.76	−14.84	−18.65
0.80	−19.53	53.91	−16.65	14.89	−47.83	7.30	−1.98	−9.88	−24.54	−13.35	
0.90	−24.29	66.30	−21.83	19.39	−53.31	9.00	−2.35	−7.09	−21.29	−9.52	
1.00	−30.28	82.02	−27.82	24.47	−57.90	10.50	−2.74	−1.74	−16.68	−4.23	−11.28
1.10	−38.15	102.45	−36.57	31.74	−63.12	12.14	−3.05	5.44	−10.23	3.00	
1.25	−54.38	144.78	−55.84	47.02	−73.99	15.10	−3.02	19.67	4.29	18.90	10.52
1.50	−99.70	262.47	−114.98	89.25	−97.96	21.35	2.18	62.61	48.95	66.59	58.43
1.75	−181.44	476.85	−234.40	156.91	−133.36	28.20	29.79	142.55	137.56	159.39	
2.00	−316.84	846.93	−447.64	230.62	−182.68	31.32	121.97	283.67	298.85	326.50	
6.00	−7.66	24.68	−5.78	5.24	−15.93	2.47	−1.78	1.25	−0.38	2.45	
6.50	−2.00	5.64	−0.76	0.69	−6.99	0.69	−0.34	−3.07	−3.62	−2.23	
7.00	−0.02	1.23	−0.14	0.12	−2.99	0.18	−0.02	−1.64	−2.51	−1.83	
8.00	0.06	0.05	0.00	0.00	−0.63	0.01	−0.09	−0.60	−0.77	−0.60	

[a] The total DFT-SAPT energy, as well as MP2, SCS-MP2, and DFT+Disp interaction energies are also given. Energy values in kilojoules per mole; distances in angstroms. Note that for distances $r > 3.5$ Å the complex is *exohedral*.

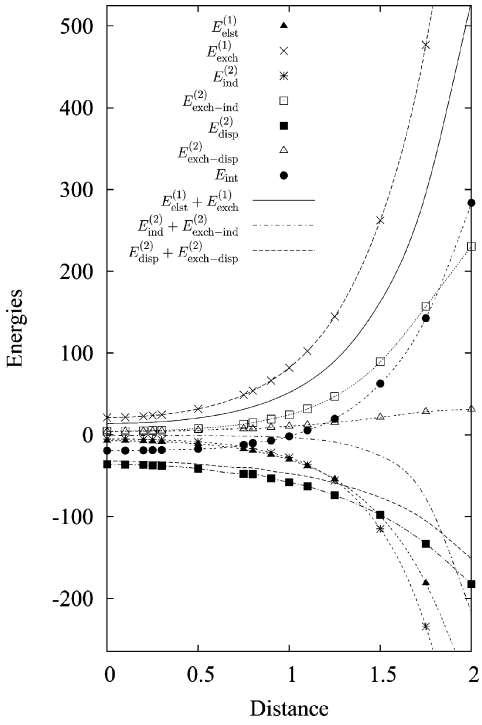


Figure 3. Components of the interaction energy for the PP orientation of $H_2@C_{60}$. Energy values are in kilojoules per mole, distances are in angstroms.

the opposite sides from the center of C_{60}, at a distance $r =$ 1.5 Å each, they will exhibit a strong repulsion from the cage wall, but the hydrogen molecules will not repel each other. Therefore, the most interesting geometries for the $2H_2@C_{60}$ complex are those with hydrogen molecules at distances from 1.0 to 1.5 Å from the C_{60} center. In this region, a minimum of the $2H_2@C_{60}$ interaction energy should be expected.

Accuracy of the Present Calculations. The DFT-SAPT approach is far too expensive to perform a geometry optimization of the $H_2@C_{60}$ and $2H_2@C_{60}$ complexes, raising questions about the accuracy of our results. The problem is especially important in the latter case, in which strain should lead to the bond-length distortion. To address this issue, we performed several additional test calculations.

The most important question to be answered is: Does an appropriate deformation of the host and/or guest allow the insertion of a second H_2 molecule into C_{60}? As already noted, the optimization of the geometry is out of the question in our case, and therefore we tackled this problem in another way. We unphysically enlarged the fullerene cage by increasing all carbon−carbon distances by 5% and calculated the interaction energy for the TP orientation and distances $r = 1.1$ and 1.25 Å. The resulting DFT-SAPT interaction energies are equal to −1.52 and +4.85 kJ/mol, respectively. This simple test shows that the unphysically large blowup of the cage shifts the zero point of PES from about 1.0 to ca. 1.2 Å, the value still too small to avoid a repulsion between two hydrogen molecules. Therefore, it seems highly improbable that much smaller changes in the geometry of C_{60} during the geometry optimization would allow a deformed $H_2@C_{60}$ to accept one more hydrogen molecule. Additionally, we found that a change in the distance between the hydrogen atoms ($R_{H-H} = 0.7408$ Å) has a negligible effect of 0.1 kJ/mol on the DFT-SAPT energy for the PP orientation at $r = 1.0$ Å. Summarizing, these data strongly indicate that neither a deformation of the host nor that of the guest would result in stabilizing of the complex of C_{60} with two hydrogen molecules.

Finally, the basis set effects were analyzed by performing the DFT-SAPT calculations in a sequence of DZP, TZVP, and TZVPP basis sets for the TP orientation at $r = 0$. The results presented in Figure 4 indicate that the quality of the dispersion energy depends crucially on the basis set used, while all other SAPT corrections are almost saturated even for the smallest DZP basis set. However, because of the importance of the dispersion energy the DZP basis set cannot be used for the $H_2@C_{60}$ complex, as it recovers only 61% of the TZVPP dispersion term. On the other hand, the TZVP $E^{(2)}_{disp}$ energy is much closer to the TZVPP value (its absolute

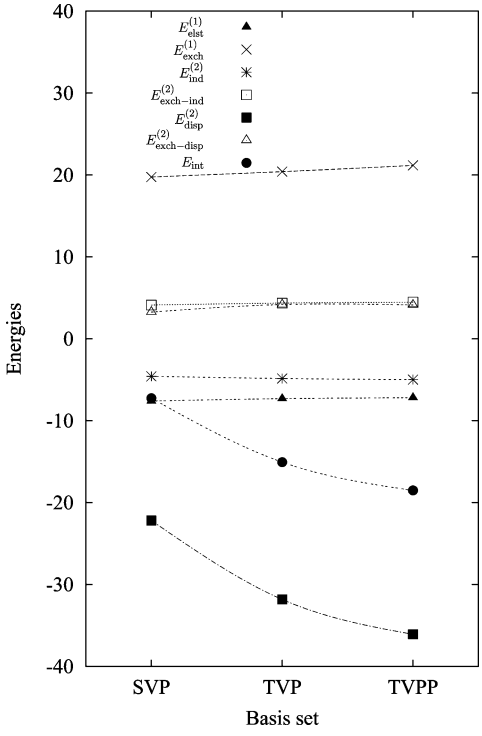


Figure 4. Basis set dependence of the SAPT components for the TP orientation of $H_2@C_{60}$ at distance $r = 0$. Energies are in kilojoules per mole.

value is smaller by 12%), allowing us to draw a conclusion that the presented results should be reliable. The remaining error resulting from the unsaturation of the basis can be conservatively estimated as about 15%.

C_{60} with Two H_2 Molecules. The results of the calculations of the interaction energy of fullerene and two hydrogen molecules are presented in Table 3. In this table, in addition to the total interaction energy E_{int}^{hybrid}, also the following quantities are presented: a sum of the DFT-SAPT interaction energies of the first and second hydrogen molecule with the fullerene, the DFT-SAPT interaction energy for the dimer of the hydrogen molecule, the MP2 nonadditive contribution, and the nonadditive CKS third-order dispersion term. The total supermolecular MP2 interaction energy is also listed for comparison.

An examination of Table 3 shows that three-body effects for this case are rather small (less than 10% of the total interaction energy). Usually MP2 and third-order dispersion nonadditive effects are of the opposite signs, which makes the total nonadditive contribution even smaller. The third-order dispersion energy is found to be insignificant for almost all tested geometries. However, the MP2 supermolecular method cannot be used for the $2H_2@C_{60}$ case, anyway, since the two-body energies predicted by MP2 are too attractive for $H_2@C_{60}$ in comparison to the DFT-SAPT reference values.

The shape of the $2H_2@C_{60}$ potential is determined by the two-body effects. The anisotropy of the three-body interaction energy is quite pronounced. In all tested cases, the interaction energy is positive, denoting that the endohedral complex of fullerene with two H_2 molecules is not stabilized. The minimum repulsion (ca. 24.7 /kJ/mol) occurs for both

Table 3. Components of the Interaction Energy for Various Orientations of Hydrogen Molecules in the $2H_2@C_{60}$ Complex[a]

r	$\sum E_{int}^{SAPT}$ ($H_2@C_{60}$)[b]	E_{int}^{SAPT} ($H_2\cdots H_2$)	E_{int}^{MP2}[3,3]	$E_{disp}^{(3)}$(CKS)[3,3]	E_{int}^{hybrid}	E_{int}^{MP2}
			TP			
0.50	−33.92	537.66	−1.04	−6.34	496.35	583.51
0.75	−23.07	133.97	4.46	−2.08	113.27	104.53
0.80	−18.95	100.56	4.69	−1.66	84.64	68.14
0.90	−11.50	55.22	4.63	−1.07	47.28	23.51
1.00	−0.04	29.18	4.19	−0.58	32.75	0.45
1.10	15.33	14.64	3.63	−0.25	33.35	2.97
1.25	46.86	4.77	2.85	−0.17	54.31	23.92
1.50	138.60	0.36	2.10	−0.03	141.03	116.52
1.75	305.94	−0.19	1.94	−0.05	307.65	305.65
			XP			
0.50	−34.51	469.32	0.59	−25.23	392.71	490.25
0.75	−24.58	120.33	4.47	−3.42	96.81	85.43
0.90	−14.18	48.97	4.08	−0.55	38.31	13.19
1.00	−3.49	25.98	3.59	−1.18	24.91	−3.16
1.10	10.87	13.44	3.12	−0.91	26.53	−4.11
1.25	39.33	4.81	2.56	−0.22	46.49	15.32
1.50	125.22	0.71	2.10	−0.18	127.85	100.21
			SP			
0.50	−33.56	460.05	0.77	−7.99	419.27	496.42
0.75	−22.92	132.00	4.32	−2.45	110.95	100.79
1.00	−0.50	29.37	3.86	−0.88	31.85	4.10
1.25	46.23	4.97	2.61	−2.63	51.18	23.19
1.50	136.73	0.44	1.97	−0.15	138.98	115.25
			XH			
0.50	−34.16	476.35	0.48	−7.08	435.59	497.80
0.75	−24.42	121.31	4.39	−2.80	98.48	86.46
0.90	−13.82	49.34	3.99	−1.70	37.82	13.56
1.00	−3.98	26.19	3.49	−0.91	24.79	−3.07
1.10	9.95	13.55	3.00	−0.59	25.91	−4.40
1.25	38.85	4.85	2.41	−0.46	45.66	13.86
1.50	119.77	0.73	1.89	−0.24	122.14	92.90

[a] The total interaction energy E_{int}^{hybrid} is a sum of additive DFT-SAPT energies and nonadditive (E_{int}^{MP2}[3,3] and $E_{disp}^{(3)}$(CKS)[3,3]) energies; see eq 6. The total supermolecular MP2 interaction energy is listed in the last column. Energy values in kilojoules per mole; distances in angstroms. [b] A sum of interaction energies of both fullerene−hydrogen molecule pairs; see eq 4.

"crossed" structures for hydrogen molecules at distance of 2.0 Å from each other and of 1.0 Å from the center of C_{60}. It is noteworthy that these two orientations are different from the global-minimum orientation of the H_2 dimer (corresponding to the TP structure).[85] Evidently, the TP orientation is more repulsive (ca. 32.8 kJ/mol) since in such an orientation one hydrogen atom (of the H_2 molecule perpendicular to a pentagon ring) "touches" the cage wall sooner than in the case of the parallel orientation. Thus, for the "crossed" structures, the minimum is a result of an interplay of the two-body interaction energies of the $H_2\cdots H_2$ and $H_2@C_{60}$ species. It seems unlikely that interaction energies of other orientations would be significantly lower than the tried ones. Therefore, one can conclude that the present method does not yield the stabilization of the $2H_2@C_{60}$ complex. It can also be observed that E_{int}^{MP2} predicts falsely a small stabilization effect for "crossed" structures, which can be explained by too attractive interaction energies predicted for $H_2@C_{60}$ (Table 2 and the Supporting Information).

In view of the recent synthesis of two hydrogen molecules in a closed C_{70} cage,[18] we performed the DFT-SAPT calculations for several points of PES for the $2H_2@C_{70}$ complex. Because of the limitations of our third-order

dispersion code, only the additive part of the interaction energy was obtained. The geometry of the C_{70} fullerene was taken from ref 88. The X1 structure of the H_2 dimer was used with the geometrical centers of H_2 lying on the fivefold symmetry axis of C_{70}, on the opposite sides from the cage center at distances $r = 1.2$, 1.3, 1.4, and 1.5 Å from this center. The asymptotic shift of the bulk xc potential of the C_{70} fullerene was taken as 0.0596 hartree. A smaller TZVP basis was used. The additive part of the DFT-SAPT energy for these distances is equal to −16.3, −19.8, −13.6, and −9.1 kJ/mol, respectively. It can be noted that the largest (in absolute value) interaction energy still occurs for the repulsive geometry of the $H_2\cdots H_2$ dimer. Since the result is obtained in the TZVP basis and the attractive dispersion energy benefits the most from using the larger TZVPP basis, it can be estimated that the value of the interaction energy can be about 10−20% lower in the full basis set limit. The experience gained from the $2H_2@C_{60}$ case allows one to estimate the possible nonadditive effects as at most 10% of the total interaction energy. Thus, in agreement with the experimental findings,[11,18] the DFT-SAPT approach yields the stabilization of two hydrogen molecules inserted into the C_{70} fullerene and the destabilization of the smaller $2H_2@C_{60}$ complex.

Summary and Conclusions

The highly accurate DFT-SAPT method with density fitting used for two-electron repulsion integrals was shown to be applicable for an analysis of selected points of the potential energy surface for the nonbonding interactions of the C_{60} fullerene with hydrogen molecules.

The calculations were performed with DFT-SAPT in a reasonably large TZVPP orbital basis for selected orientations of one and two H_2 molecules inside the C_{60} fullerene. The nonadditive effects were modeled by a recently proposed hybrid method.[73] For the endohedral complex $H_2@C_{60}$, a small stabilization effect of about 19.4 kJ/mol (4.6 kcal/mol) was found, with the minimum of PES in the center of the fullerene. It can be noted that this value agrees nicely with a recent estimate of Slanina et al.,[47] who predicted the stabilization of at least 4 kcal/mol for this species. The PES of $H_2@C_{60}$ is almost flat in the vicinity of the cage center and nearly isotropic, especially in the attractive region. This result is consistent with a recent theoretical study of the translation-rotation spectrum of H_2 confined in C_{60},[89] where the first rotational level of $H_2@C_{60}$ is virtually identical to the level for the free hydrogen molecule. The hydrogen molecule inside the fullerene is bound mainly by the dispersion interaction, while the first-order exchange term represents the main repulsive component of the interaction energy. However, other SAPT corrections are far from being negligible. For instance, the induction energy is of the same order of magnitude as the dispersion energy, but is strongly quenched by its exchange counterpart in the vicinity of the cage center. A small exohedral minimum, expected on the basis of model considerations, was also observed.

For the $2H_2@C_{60}$ complex, no stabilization effect was found. This finding is in agreement with the lack of the experimental reports of two H_2 molecules inside the opened and closed C_{60} cage and with only a small amount of the $2H_2@C_{70}$ obtained in the mixture with $H_2@C_{70}$. The lowest repulsion for the $2H_2@C_{60}$ complex occurs for the "crossed" orientation of the hydrogen molecules, which are separated by ca. 2.0 Å from each other. Interestingly enough, the stabilization of $H_2@C_{60}$ and destabilization of $2H_2@C_{60}$ was also predicted by a simple MM model.[16,19] For the same orientation of the hydrogen molecules in a larger C_{70} fullerene, separated by 2.6 Å, the DFT-SAPT method yields the negative interaction energy, confirming, in agreement with recent experimental findings, the stability of $2H_2@C_{70}$. Interestingly, also in this case the MM method yielded the stabilization of both $H_2@C_{70}$ and $2H_2@C_{70}$ species, correctly predicting their energy difference.[19]

Acknowledgment. Computations were carried out using the mixed Woodcrest/Opteron cluster at the Lehrstuhl für Theoretische Chemie Erlangen and the Opteron cluster at the Computer Center of the Faculty of Chemistry, University of Warsaw. A.H. gratefully acknowledges the funding of the German Research Council (DFG), through the Cluster of Excellence "Engineering of Advanced Materials" (www.eam.uni-erlangen.de).

Supporting Information Available: Results of calculations of DFT-SAPT, MP2, SCS-MP2, and DFT-Disp interaction energies, as well as DFT-SAPT energy components for the TP, TH, SP, and HP orientations of the $H_2@C_{60}$ complex. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Kroto, H. W.; Heath, J. R.; O'Brien, S. C.; Curl, R. F.; Smalley, R. E. *Nature (London)* **1985**, *318*, 162. (b) Kroto, H. *Angew. Chem., Int. Ed.* **1997**, *36*, 1579.

(2) (a) Krätschmer, W.; Fostiropoulos, K.; Huffman, D. *Chem. Phys. Lett.* **1990**, *170*, 167. (b) Krätschmer, W.; Lamb, L. D.; Fostiropoulos, K.; Huffman, D. R. *Nature (London)* **1990**, *347*, 354.

(3) Stoddart, J. F. *Angew. Chem., Int. Ed.* **1991**, *30*, 70.

(4) (a) Töth, E.; Bolskar, R. D.; Borel, A.; González, G.; Helm, L.; Merbach, A. E.; Sitharaman, B.; Wilson, L. J. *J. Am. Chem. Soc.* **205**, *127*, 799. (b) Bolskar, R. D. *Nanomedicine* **2008**, *3*, 201. (c) MacFarland, D. K.; Walker, K. L.; Lenk, R. P.; Wilson, S. R.; Kumar, K.; Kepley, C. L.; Garbow, J. R. *J. Med. Chem.* **2008**, *51*, 3681. (d) Watanabe, K.; Ishioka, N. S.; Sekine, T.; Kudo, H.; Shimomura, H.; Muratsu, H.; Kume, T. *J. Radioanal. Nucl. Chem.* **2005**, *266*, 499. (e) Wilson, L. J.; Cagle, D. W.; Thrash, T. P.; Kennel, S. J.; Mirzadeh, S.; Alford, J. M.; Ehrhardt, G. J. *Coord. Chem. Rev.* **1999**, *192*, 199.

(5) Mauter, M. S.; Elimelech, M. *Environ. Sci. Technol.* **2008**, *42*, 5843.

(6) Fernandez, G.; Sanchez, L.; Perez, E. M.; Martin, N. *J. Am. Chem. Soc.* **2008**, *130*, 10674.

(7) Stefan-van Staden, R. I.; Lal, B. *Anal. Lett.* **2006**, *39*, 1311.

(8) (a) Shin, W. H.; Yang, S. H.; Goddard, W. A.; Kang, J. K. *Appl. Phys. Lett.* **2006**, *88*, 53111. (b) Denis, P. A. *J. Phys. Chem. C* **2008**, *112*, 2791.

(9) Pupysheva, O. V.; Farajian, A. A.; Yakobson, B. I. *Nano Lett.* **2008**, *8*, 767.

(10) Bénard, P.; Chagine, R.; Chandonia, P. A.; Cossement, D.; Dorval-Douville, G.; Lafi, L.; Lachance, P.; Paggiaro, R.; Poirier, E. *J. Alloys Compd.* **2007**, *446*, 380.

(11) Komatsu, K.; Murata, M.; Murata, Y. *Science* **2005**, *307*, 238.

(12) (a) Rubin, Y. *Chem.−Eur. J.* **1997**, *3*, 1009. (b) Rubin, Y. *Chimia* **1998**, *52*, 118. (c) Murata, Y.; Murata, M.; Komatsu, K. *J. Am. Chem. Soc.* **2003**, *125*, 7152. (d) Sawa, H.; Wakabayashi, Y.; Murata, Y.; Murata, M.; Komatsu, K. *Angew. Chem., Int. Ed.* **2005**, *44*, 1981.

(13) Rubin, Y. *Top. Curr. Chem.* **1999**, *199*, 97.

(14) Murata, M.; Murata, Y.; Komatsu, K. *Chem. Commun.* **2008**, 6083.

(15) Patchkovskii, S.; Thiel, W. *J. Am. Chem. Soc.* **1996**, *118*, 7164.

(16) Dodziuk, H.; Dolgonos, G.; Lukin, O. *Carbon* **2001**, *39*, 1907.

(17) Murata, Y.; Maeda, S.; Murata, M.; Komatsu, K. *J. Am. Chem. Soc.* **2008**, *130*, 6702.

(18) Murata, M.; Maeda, S.; Morinaka, Y.; Murata, Y.; Komatsu, K. *J. Am. Chem. Soc.* **2008**, *130*, 15800.

(19) Dodziuk, H. *Chem. Phys. Lett.* **2005**, *410*, 39.

(20) Dresselhaus, M. S.; Dresselhaus, G.; Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes*; Academic Press: San Diego, CA, 1996; pp 60−79.

(21) Seifert, G. *Solid State Ionics* **2004**, *168*, 265.

(22) Koi, N.; Oku, T. *Sci. Technol. Adv. Mater.* **2004**, *5*, 625.

(23) Ramachandran, C. N.; Roy, D.; Sathyamurthy, N. *Chem. Phys. Lett.* **2008**, *461*, 87.

(24) Osawa, E.; Musso, H. *Angew. Chem., Int. Ed.* **1983**, *22*, 1.

(25) Dannenberg, J. J. *J. Mol. Struct.: THEOCHEM* **1997**, *401*, 279.

(26) Turker, L.; Erkoc, S. *Chem. Phys. Lett.* **2006**, *426*, 222.

(27) Ren, Y. X.; Ng, T. Y.; Liew, K. M. *Carbon* **2006**, *44*, 397.

(28) Dolgonos, G. *J. Mol. Struct.: THEOCHEM* **2005**, *732*, 239.

(29) Dodziuk, H. *Chem. Phys. Lett.* **2006**, *426*, 224.

(30) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159.

(31) (a) Scuseria, G. E. Theoretical Studies of Fullerenes. In *Modern Electronic Structure Theory*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; pp 279−310. (b) Scuseria, G. E. *Science* **1996**, *271*, 942.

(32) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

(33) Kowalski, K.; Hammond, J. R.; de Jong, W. A.; Sadlej, A. J. *J. Chem. Phys.* **2008**, *129*, 226101.

(34) Jeziorska, M.; Jeziorski, B.; Cizek, J. *Int. J. Quantum Chem.* **1987**, *32*, 149.

(35) Cioslowski, J. *J. Am. Chem. Soc.* **1991**, *113*, 4139.

(36) (a) Bartlett, R. J.; Lotrich, V. F.; Schweigert, I. V. *J. Chem. Phys.* **2005**, *123*, 62205. (b) Kamiya, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 6010. (c) Hobza, P.; Zahradnik, R.; Müller-Dethlefs, K. *Collect. Czech. Chem. Commun.* **2006**, *71*, 443.

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

(38) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: New York, 2000; p 236.

(39) Bartlett, R. J.; Lotrich, V. F.; Schweigert, I. V. *J. Chem. Phys.* **2005**, *123*, 062205.

(40) Bartlett, R. J.; Grabowski, I.; Hirata, S.; Ivanov, S. *J. Chem. Phys.* **2005**, *122*, 034104.

(41) Furche, F.; Van Voorhis, T. *J. Chem. Phys.* **2005**, *122*, 164106.

(42) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

(43) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(44) Elstner, M.; Frauenheim, T.; Kaxiras, E.; Seifert, G.; Suhai, S. *Phys. Status Solidi B* **2000**, *217*, 357.

(45) Ganji, M. D.; Zare, K. *Mol. Simul.* **2008**, *34*, 821.

(46) (a) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554. (b) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(47) Slanina, Z.; Pulay, P.; Nagase, S. *J. Chem. Theory Comput.* **2006**, *2*, 782.

(48) (a) Peres, T.; Cao, B. P.; Cui, W. D.; Lifshitz, C.; Khong, A.; Cross, R. J.; Saunders, M. *Int. J. Mass Spectrom.* **2001**, 210–241. (b) Suetsuna, T.; Dragoe, N.; Harneit, W.; Weidinger, A.; Shimotani, H.; Ito, S.; Takagi, H.; Kitazawa, K. *Chem.−Eur. J.* **2002**, *8*, 5079. (c) Suetsuna, T.; Dragoe, N.; Harneit, W.; Weidinger, A.; Shimotani, H.; Ito, S.; Takagi, H.; Kitazawa, K. *Chem.−Eur. J.* **2002**, *9*, 598.

(49) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(50) Hobza, P.; Selzle, H. L.; Schlag, H. W. *J. Phys. Chem.* **1996**, *100*, 18790.

(51) Yang, C.-K. *Carbon* **2007**, *45*, 2451.

(52) Lee, T. B.; McKee, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 17610.

(53) Chuang, S.-C.; Murata, Y.; Murata, M.; Komatsu, K. *J. Org. Chem.* **2007**, *72*, 6447.

(54) Dolgonos, G. *Carbon* **2008**, *46*, 704.

(55) Yang, C.-K. *Carbon* **2008**, *46*, 705.

(56) Hu, Y.; Ruckenstein, E. *J. Chem. Phys.* **2003**, *119*, 10073.

(57) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.

(58) Szalewicz, K.; Patkowski, K.; Jeziorski, B. *Struct. Bonding (Berlin)* **2005**, *116*, 43.

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

(60) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorska, M.; Jeziorski, B.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszyński, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wheatley, R. J.; Wormer, P. E. S.; Zuchowski, P. S. SAPT2008: An Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies; University of Delaware and University of Warsaw: Newark, DE and Warsaw, Poland, 2008. http://www.physics.udel.edu/~szalewic/SAPT/SAPT.html.

(61) Moszynski, R.; Heijmen, T. G. A.; Jeziorski, B. *Mol. Phys.* **1996**, *88*, 741.

(62) (a) Jeziorski, B.; Moszynski, R.; Rybak, S.; Szalewicz, K. In *Many-Body Methods in Quantum Chemistry*; Kaldor, U., Ed.; Lecture Notes in Chemistry 52; Springer: New York, 1989; p 65. (b) Rybak, S.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1991**, *95*, 6576. (c) Moszynski, R.; Jeziorski,

B.; Ratkiewicz, A.; Rybak, S. *J. Chem. Phys.* **1993**, *99*, 8856. (d) Moszynski, R.; Jeziorski, B.; Szalewicz, K. *Int. J. Quantum Chem.* **1993**, *45*, 409. (e) Moszynski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1994**, *100*, 1312. (f) Moszynski, R.; Cybulski, S. M.; Chalasinski, G. *J. Chem. Phys.* **1994**, *100*, 4998.

(63) (a) Moszynski, R.; Jeziorski, B.; Rybak, S.; Szalewicz, K.; Williams, H. L. *J. Chem. Phys.* **1994**, *100*, 5080. (b) Williams, H. L.; Szalewicz, K.; Moszynski, R.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 8058.

(64) (a) Korona, T.; Jeziorski, B. *J. Chem. Phys.* **2006**, *125*, 184109. (b) Korona, T. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6004. (c) Korona, T.; Jeziorski, B. *J. Chem. Phys.* **2008**, *128*, 144107. (d) Korona, T. *J. Chem. Phys.* **2008**, *122*, 224104. (e) Korona, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6509.

(65) Hesselmann, A.; Jansen, G. *J. Chem. Phys.* **2000**, *112*, 6949.

(66) (a) Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646. (b) Jansen, G.; Hesselmann, A. *J. Phys. Chem. A* **2001**, *105*, 11156.

(67) (a) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464. (b) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319. (c) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778.

(68) Hesselmann, A.; Jansen, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010.

(69) (a) Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301. (b) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(70) (a) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103. (b) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730. (c) Podeszwa, R.; Szalewicz, K. *Chem. Phys. Lett.* **2005**, *412*, 488. (d) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Chem. Theory Comput.* **2006**, *2*, 400.

(71) Tekin, A.; Jansen, G. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1680.

(72) (a) Moszynski, R.; Wormer, P. E. S.; Jeziorski, B.; van der Avoird, A. *J. Chem. Phys.* **1995**, *103*, 8058. (b) Moszynski, R.; Wormer, P. E. S.; Jeziorski, B.; van der Avoird, A. *J. Chem. Phys.* **1997**, *107*, 672. (c) Lotrich, V. F.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 9668. (d) Wormer, P. E. S.; Moszynski, R.; van der Avoird, A. *J. Chem. Phys.* **2000**, *112*, 3159. (e) Lotrich, V. F.; Szalewicz, K. *J. Chem. Phys.* **2000**, *112*, 112.

(73) Podeszwa, R.; Szalewicz, K. *J. Chem. Phys.* **2007**, *126*, 194101.

(74) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. MOLPRO, A Package of ab Initio Programs, version 2008.2; Cardiff University: Cardiff, UK, 2008. http://www.molpro.net.

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

(76) Grüning, M.; Gritsenko, O. V.; van Gisergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 652.

(77) Lichtenberger, D. L.; Nebesny, K. W.; Ray, C. D.; Huffman, D. R.; Lamb, L. D. *Chem. Phys. Lett.* **1991**, *176*, 203.

(78) Computational Chemistry Comparison and Benchmark DataBase. http://cccbdb.nist.gov (Accessed July 1, 2008).

(79) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

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

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

(82) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.

(83) Hedberg, K.; Hedberg, L.; Bethune, D.; Brown, C. A.; Dorn, H. C.; Johnson, R. D.; de Vries, M. *Science* **1991**, *254*, 410.

(84) Jeziorska, M.; Jankowski, P.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **2000**, *113*, 2957.

(85) Hinde, R. *J. Chem. Phys.* **2008**, *128*, 154308.

(86) Bemish, R. J.; Oudejans, L.; Miller, R. E.; Moszynski, R.; Heijmen, T. G. A.; Korona, T.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **1998**, *109*, 8968.

(87) Olthof, E. H. T.; van der Avoird, A.; Wormer, P. E. S. *J. Chem. Phys.* **1995**, *104*, 832.

(88) Download Cartesian Coordinates. http://www.cochem2.tutkie.tut.ac.jp/Fuller/higher/higherE.html (Accessed December 23, 2008).

(89) Xu, M.; Sebastianelli, F.; Bacic, Z.; Lawler, R.; Turrob, N. J. *J. Chem. Phys.* **2008**, *128*, 011101.

CT900108F

JCTC Journal of Chemical Theory and Computation

Exploring the Reactivity Trends in the E2 and S_N2 Reactions of $X^- + CH_3CH_2Cl$ (X = F, Cl, Br, HO, HS, HSe, NH_2 PH_2, AsH_2, CH_3, SiH_3, and GeH_3)

Xiao-Peng Wu,[†,‡] Xiao-Ming Sun,[‡] Xi-Guang Wei,[‡] Yi Ren,*[,‡] Ning-Bew Wong,*[,†] and Wai-Kee Li[§]

Department of Biology and Chemistry, City University of Hong Kong, Kowloon, Hong Kong, College of Chemistry, Key Laboratory of Green Chemistry and Technology, Ministry of Education, and Key State Laboratory of Biotherapy, Sichuan University, Chengdu 610064, People's Republic of China, and Department of Chemistry, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Received January 22, 2009

Abstract: The reactivity order of 12 anions toward ethyl chloride has been investigated by using the G2(+) method, and the competitive E2 and S_N2 reactions are discussed and compared. The reactions studied are $X^- + CH_3CH_2Cl \rightarrow HX + CH_2{=}CH_2 + Cl^-$ and $X^- + CH_3CH_2Cl \rightarrow CH_3CH_2X + Cl^-$, with X = F, Cl, Br, HO, HS, HSe, NH_2 PH_2, AsH_2, CH_3, SiH_3, and GeH_3. Our results indicate that there is no general and straightforward relationship between the overall barriers and the proton affinity (PA) of X^-; instead, discernible linear correlations only exist for the X's within the same group of the periodic table. Similar correlations are also found with the electronegativity of central atoms in X, deformation energy of the E2 transition state (TS), and the overall enthalpy of reaction. It is revealed that the electronegativity will significantly affect the barrier height, and a more electronegative X will stabilize the E2 and S_N2 transition states. Multiple linear regression analysis shows that there is a reasonable linear correlation between E2 (or S_N2) overall barriers and the linear combination of PA of X^- and electronegativity of the central atom.

1. Introduction

Base-induced bimolecular elimination (E2) and bimolecular nucleophilic substitution (S_N2) reactions are two fundamental organic reactions in synthesis. They play an important role in the development of modern mechanistic physical organic chemistry.[1] In many cases, E2 and S_N2 pathways are usually competitive processes. The S_N2/E2 competition in the gas phase and condensed phase has been exhaustively investigated experimentally[2−12] and theoretically[13−20] over the past 30 years, which helps us with a better understanding of the factors controlling the competition between them. For example, by direct detection of the neutral products, Jones et al.[2] found that elimination was preferred for the gas-phase reactions between the methoxide ion (CH_3O^-) and 1-bromopropane ($CH_3CH_2CH_2Br$). Their results in the gas phase contrast sharply with those in solution studies,[3] which show an overwhelming preference for S_N2. Although the neutral detection methods could provide useful information, their applications are restricted by experimental difficulties and limitations. Later, Gronert el al.[5−7] proposed a novel approach for analyzing the product mixtures to investigate the gas-phase S_N2/E2 competition. By using a double charged nucleophile (Nu) where the anionic site is nucleophilic and the other is unreactive, its reaction with alkyl halide will produce two charged species: a halide ion and an alkylated (S_N2 pathway) or a protonated (E2 pathway) nucleophile. In this way, two mechanisms can be identified. The reactions

* Corresponding author fax: +86-28-85412907 (Y.R.), +852-27887406 (N.-B.W.); e-mail: yiren57@hotmail.com (Y.R.), bhnbwong@cityu.edu.hk (N.-B.W.).

[†] University of Hong Kong.

[‡] Sichuan University.

[§] The Chinese University of Hong Kong.

10.1021/ct900041y CCC: $40.75
Published on Web 04/30/2009

of dialkyl ethers with bases have been the subject of several studies. In the early works by DePuy and Bierbaum,[8,9] the flowing afterglow (FA) technique was employed to study the gas-phase reactions of a series of dialkyl ethers with amide and hydroxide ions. It was observed that cyclic and acyclic ethers with β-hydrogens react rapidly in the gas phase with both NH_2^- and OH^- by elimination rather than substitution pathways due to ring-strain release and reaction exothermicity. In a different study,[10] DePuy et al. investigated the gas-phase E2/S_N2 competition by measuring the rate coefficients for the gas-phase reactions of alkyl chlorides and bromides with a set of nucleophiles, F^-, Cl^-, RO^- (R = H, CH_3, CF_3CH_2, $C_2F_5CH_2$), and RS^- (R = H and NH_2). On the basis of their obtained reactivity trends, it was found that Nus (F^- and RO^-) involving first-row elements are capable of undergoing both substitution and elimination, whereas the second-row Nus (e.g., HS^- and H_2NS^-) are mainly limited to substitution reactions. Moreover, RS^- induces elimination much less readily than does the RO^- even when the two anions have identical basicities.

In an early theoretical study, Yamabe et al.[13] studied the gas-phase E2 and S_N2 reactions of fluoride anion with fluoroethane by ab initio calculations at the level of HF/3-21G(+p). Comparison of the two competitive reaction pathways reveals that the mechanism of E2 reaction and the geometry of the E2 TS are completely different from those of the S_N2 reaction. Meanwhile, their Hartree−Fock calculations showed that the activation barrier of the E2 reaction is higher than that of the S_N2 reaction, which is in disagreement with the experimental results of Ridge and Beauchamp[12] on the $F^- + CH_3CH_2F$ system, where the E2 pathway is more favored in the gas phase. More recently, Bickelhaupt et al.[14] made an ab initio and DFT benchmark study on the E2 and S_N2 reactions of $X^- + CH_3CH_2X$ (X = F, Cl), indicating that the *anti*-E2 pathway dominates for $F^- + CH_3CH_2F$, and the backside S_N2 pathway is more favorable for $Cl^- + CH_3CH_2Cl$, while *syn*-E2 is the least favorable pathway in all cases, indicating that a fairly high level of theory is required in the studies on the S_N2 and E2 reactions. Gronert et al.[15−20] carried out a series of comprehensive theoretical studies on elimination reactions as well as S_N2 pathways with ab initio calculations. Using the G2+ approach, Gronert and co-workers[15] also studied the reaction of F^- with CH_3CH_2F and concluded that the E2 should dominate because its barrier is smaller and its pathway is less demanding entropically. At the level of MP4SDQ/6-31+G(d, p)/HF/6-31(+)G(d), Gronert et al.[16] investigated the reactions of F^- and PH_2^- with CH_3CH_2C1 and discussed the competition between S_N2 and E2 mechanisms for the first- and second-row nucleophiles. Their theoretical results indicated that the first-row Nus are well-suited for both S_N2 and E2 reactions, whereas second-row Nus with similar basicity are more confined to S_N2 reactions, which is consistent with the results of the aforementioned gas-phase experimental studies by DePuy et al.[10] The enhanced reactivity of fluoride anion could be rationalized by electron reorganization; that is, less electron density redistribution during either reaction will lead to a lower activation barrier. In another study, Gronert and co-workers evaluated the effect of methyl substitution on E2 and S_N2 mechanisms for the gas-phase reactions of F^- with $(CH_3)_2CHCl$ and $CH_3CH_2CH_2Cl$ at the MP2/6-31+G(d,p)//HF/6-31+G(d) level. A comparison of the activation barriers of the S_N2 and E2 reactions predicts that elimination will dominate in the reaction of propyl chloride.[17]

Recently, several studies have focused on a competitive reaction system, $ClO^- + CH_3CH_2Cl$ via E2 and S_N2 channels. By dual-level generalized transition state theory and statistical calculations based on high-level correlated electronic structure calculations using MP2 theory level and modified aug-cc-pVDZ basis set (MP2/ADZP), Hu and Truhlar[21] quantitatively evaluated the rate constants and deuterium kinetic isotope effects (KIEs) for the competing S_N2/E2 reactions of ClO^- with C_2H_5Cl or C_2D_5Cl in the gas phase. It was predicted that KIEs at room temperature were "normal" ($k_H/k_D = 3.1$) for the E2 reaction but "inverse" ($k_H/k_D = 0.6$) for the S_N2 reaction. Villano et al.[22] measured the overall reaction rate constants and KIEs for the gas-phase reactions of $RCl + ClO^-$ (R = CH_3, C_2H_5, iso-C_3H_7, and *tert*-C_4H_9) using a tandem flowing afterglow-selected ion flow tube (FA-SIFT) instrument. The experimental reaction efficiencies (10%)[23] and the KIEs ($k_H/k_D = 0.99 \pm 0.01$) for the reaction of ClO^- with C_2H_5Cl were shown to differ from the theoretical values (28% and 2.4) by Hu et al.,[21] suggesting that the S_N2 channel is more prominent in experiment than the calculated prediction. They proposed that nonstatistical dynamics or errors in the calculation of the individual KIE or in the branching ratios of the two channels could account for the discrepancies between experiment and theory, and additional studies were suggested to describe nucleophilic substitution and elimination reactions more accurately. To amend the shortage of the theoretical studies in the condensed phase, recently Pabis et al.[24] studied the KIEs on the two alternative reactions, S_N2 and E2, between ClO^- and C_2H_5Cl in water using B3LYP and M06-2X[25,26] functionals with the standard 6-31+G(d,p) basis set and the polarizable continuum solvent model (PCM).[27] The results show that the KIEs obtained using both DFT functionals are in qualitative agreement. It is worth noticing that this ^{18}O-KIE is a good indicator of different mechanism.

A Brønsted-type plot of log k_{nuc} versus pK_a constructed for a series of related Nus is often used to describe the relationship of basicity with nucleophilic character for the generalized acid−base reactions, for example, S_N2, and also base-induced E2 reactions. As mentioned above, Gronert et al.[16] pointed out that there might be significantly different reactivity for the nucleophiles (or base) with similar basicity, implying that the linear Brønsted-type plot does not hold for all cases and is only valid for selective Nus. There are several theoretical studies treating the reactivity order of Nus in the S_N2 reactions. Radom et al.[28,29] reported G2(+) studies on the reactions of halide anions with methyl halides, giving the nucleophilic order of halides toward methyl halides, $F^- > Cl^- > Br^- > I^-$; Bickelhaupt et al.[30] also carried out a study on the nucleophilicity of halide anions using relativistic density functional theory (DFT) and found that the S_N2 barriers would increase along the nucleophiles F^-, Cl^-, Br^-, and I^-. Lee et al.[31] made ab initio studies on the S_N2 identity exchange reactions $RCH_2X + X^- \rightarrow X^- + RCH_2X$ for R =

CH_2CH with X = H, NH_2, OH, F, PH_2, SH, and Cl, and for R = CH_3 and CH≡C with X = Cl at the HF and MP2 levels of theory using the 6-31++G(d,p) basis set. They concluded that the activation barriers, and major structural changes, $\Delta d^{\ddagger}$(C−X), in the activation process are closely related to the electronegativity of the R and X groups (we will use the abbreviation EN for electronegativity from now on), and a stronger EN of R and/or X leads to less electronic as well as structural reorganization in the activation, which in turn would lower the energy barriers at both the HF and the MP2 levels. Uggerud,[32] using G2 calculations, investigated 18 S_N2 reactions, including $X^- + CH_3X \rightarrow XCH_3 + X^-$ and $XH + CH_3XH^+ \rightarrow {}^+HXCH_3 + XH$ (X = F, Cl, Br, OH, SH, SeH, NH_2, PH_2, and AsH_2), and analyzed the systematic periodic trends of intrinsic reactivity, finding that the barrier heights decrease on going from left to right of each row in the periodic table, and the basicity and nucleophilicity will be equivalent only in the strongly exothermic reactions.

Despite the importance of E2 reactions in organic synthesis, there has been less effort put on the reactivity in the base-induced E2 reactions than on S_N2 reactions until now. In the present work, G2(+) calculations are reported for a series of anionic E2 reactions toward ethyl chloride with 12 attacking atoms from groups 14−17 of the periodic table (eq 1). The corresponding competitive S_N2 reactions (eq 2) are also discussed for the sake of comparing the reactivity with that of the E2 reactions.

$$X^- + CH_3CH_2Cl \rightarrow CH_2{=}CH_2 + XH + Cl^- \quad (1)$$

$$X^- + CH_3CH_2Cl \rightarrow CH_3CH_2X + Cl^- \quad (X = F, Cl, Br, HO, SH, SeH, NH_2, PH_2, AsH_2, CH_3, SiH_3, \text{and } GeH_3) \quad (2)$$

In this work, our objectives are to find systematic periodic trends in reactivity for the base-induced E2 reactions and to provide a reasonable and consistent set of ab inito barrier height data. We will focus on the relationship between basicity and reactivity and make an extensive comparison between the E2 and the S_N2 reactions.

2. Computational Methods

Previous studies have proved that the high-level G2(+) theory, introduced by Radom and co-workers,[33] treated anions better with the added diffuse functions on non-hydrogen atoms and is able to provide reliable data for the anionic S_N2 and E2 reactions.[28,29,33−37] Therefore, the G2(+) theory was employed in the present study. We note here that the original G2(+) procedure corrects the zero-point vibrational energy using HF/6-31+G(d) frequencies, scaled by 0.8929. For some of the species studied here, especially some TSs, the HF/6-31+G(d) and MP2(fc)/6-31+G(d) structures are considerably different. Hence, for all of the values reported below, the zero-point energy was corrected at the MP2(fc)/6-31+G(d) level, using the recommended scaling factor of 0.98.[38] We also note here that using MP2/6-31+G(d) frequencies, in place of the HF/6-31+G(d) frequencies, has virtually no effect on the calculated proton affinities for all of the anions. Charges were calculated by the natural population analysis (NPA)[39] at the MP2(fc)/6-311+G(3df,2p) level on the MP2(fc)/6-31+G(d) geometries.

Table 1. Proton Affinity (PA) and Ethyl Cation Affinity (ECA) and Available Methyl Cation Affinity of Anions (in kJ mol^{-1})

X^-	G2(+) PA	G3 (MP2) PA[a]	exp. PA[b]	G2(+) ECA	G2 MCA[d]	exp. MCA[b]
F^-	1550.7	1553.9	1554.0	925.7	1078	1080
Cl^-	1397.5	1390.3	1395.0	791.4	950	952
Br^-	1354.5	1358.5	1353.5 ± 0.42	755.4	916	916
HO^-	1631.3	1632.2	1633.0	1000.6	1153	1159
HS^-	1472.1	1464.8	1468.0 ± 12.	869.6	1034	1033
HSe^-	1429.1	1433.4	1428.8 ± 2.9	831.9	999	
NH_2^-	1688.8	1686.2	1687.8 ±0.42	1067.5	1225	1234
PH_2^-	1539.0	1530.5	1536.0[c]	956.9	1124	1116
AsH_2^-	1500.6	1504.6	1496.0 ± 8.8	916.0	1085	
CH_3^-	1746.6	1746.0	1743.5 ± 2.9	1148.1		
SiH_3^-	1562.8	1557.3	1564.0 ± 8.8	999.8		
GeH_3^-	1517.7	1518.4	1501.0 ± 8.8	942.1		

[a] From ref 48. [b] From ref 44. [c] From ref 46. [d] From ref 32.

The geometrical characteristics of the TSs are described by the Pauling bond order, $n^{\ddagger}$, calculated according to eq 3, where r and $r^{\ddagger}$ are the bond lengths at the reactant (CH_3CH_2Cl) or the products (HX, CH_3CH_2X, and $CH_2{=}CH_2$), and at the TSs, respectively. The constant a is usually set to 0.26 or 0.3 Å. However, it has been found that a proportionality constant of a = 0.6 Å is more appropriate for the case where the bond in question has a bond order less than 1.[40−42] Based on the suggestions in the literature, a = 0.6 Å is adopted here for the calculations of bond order $n^{\ddagger}(X{-}H^{\beta})$, $n^{\ddagger}(C^{\beta}{-}H^{\beta})$, and $n^{\ddagger}(C^{\alpha}{-}Cl)$, while a = 0.3 Å is opted for $n^{\ddagger}(C^{\alpha}{-}C^{\beta})$.

$$n^{\ddagger} = \exp[(r - r^{\ddagger})/a] \quad (3)$$

Throughout this Article, all distances are in angstroms (Å) and all angles are in degrees (°), while energies are in kJ mol^{-1}. The overall barriers of the two series of reactions relative to free reactants are denoted as $\Delta H^{\ddagger}$(E2) for E2, and $\Delta H^{\ddagger}(S_N2)$ for S_N2, respectively. The MP2/6-31+G(d) optimized geometries and G2(+) energies of all reactants, products, and TSs involved in E2 (eq 1) and S_N2 reactions (eq 2) are given in the Supporting Information. The Gaussian 03 program package[43] was used in all calculations.

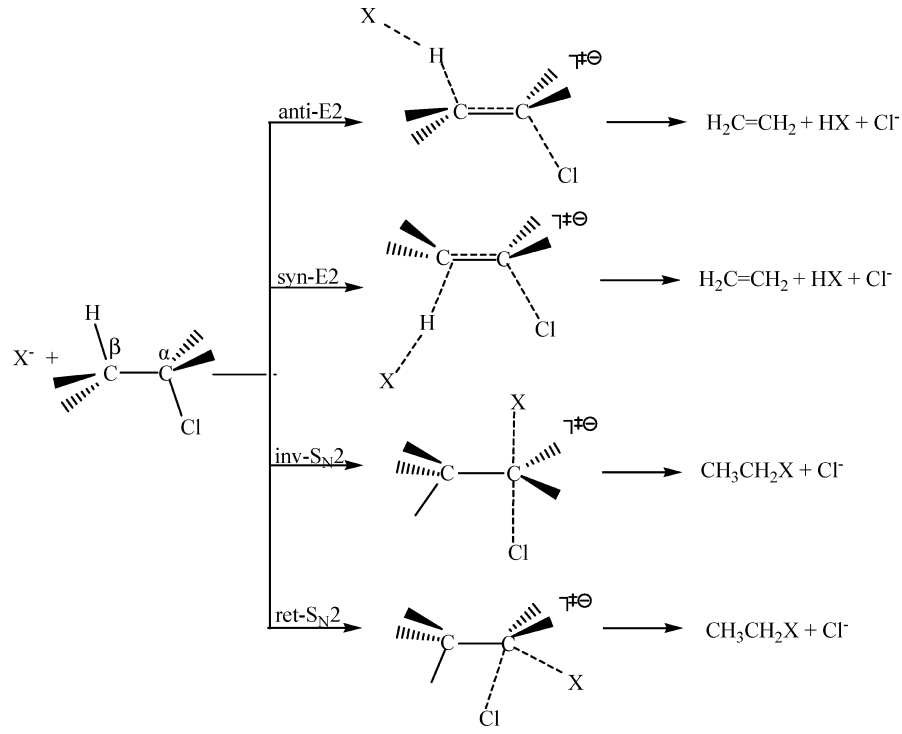
3. Results and Discussion

3.1. Proton Affinities and Ethyl Cation Affinity. The gas-phase basicity of an anion is usually measured in terms of its proton affinity (PA), that is, the negative of the enthalpy change for a gas-phase reaction like eq 4; the higher is the proton affinity, the stronger is the base and the weaker is the conjugate acid in the gas phase.

$$X^- + H^+ \rightarrow HX \quad (4)$$

The calculated and experimental PAs[44−46] of the 12 simple anionic bases given in eqs 1 and 2 are listed in Table 1. Inspection of the results in Table 1 shows that theoretical values generally compare well with their experimental counterparts, and most of them are within the so-called chemical accuracy (roughly 10 kJ mol^{-1}). The most pronounced discrepancy occurs in GeH_3^-, for which the G2(+)

Scheme 1. E2 and S_N2 Pathways for $X^- + CH_3CH_2Cl$



result is about 17 kJ mol^{-1} higher than the experimental value obtained by Decouzon et al.[45] by Fourier transform-ion cyclotron resonance (FT-ICR) spectrometry. Mayer et al.[47] and Bartmess et al.[48] also got similar results for the PA value of GeH_3^- by G2 and G3(MP2) theoretical methods, respectively.

Analogous to the methyl cation affinity (MCA) involved in the S_N2 reactions with CH_3Cl, ethyl cation affinity (ECA, defined as the enthalpy of the reaction $CH_3CH_2X \rightarrow CH_3CH_2^+ + X^-$) of the 12 anions was also calculated. Agreeing with numerous previous calculations,[49−55] the present study at the G2(+) level also demonstrates that, for $C_2H_5^+$, the structure with C_{2v} symmetry and a three-center two-electron bond with the $^1A_1$ ground electronic state is the global minimum on the potential-energy surface (PES). This result was proved recently by the highly sensitive technique of single photon IR photodissociation (IRPD) spectroscopy.[56] The calculated ECA values are found to be well correlated ($R^2 = 0.996$) with the available theoretical MCA results by the G2 method.[32]

3.2. Geometries of S_N2 and E2 Transition States. There are two possible pathways, *anti*- and *syn*-elimination, for the base-induced E2 reactions (see Scheme 1). Several previous studies[14,16] compared the energies for the *anti*- and *syn*-E2 TSs and showed that the former pathway has TSs much lower energy than the latter. For example, the *syn*-E2 TS for the F^--induced elimination of CH_3CH_2Cl lies 53.1 kJ mol^{-1} above the corresponding *anti*-E2 TS at the MP4/SDQ/6-31+G(d,p)//HF/6-31+G(d) level.[16]

For the anionic S_N2 reactions, there are also two possible pathways, back-side S_N2 with inversion of configuration and front-side S_N2 with retention of configuration (inv-, or ret-S_N2; see Scheme 1). Previous studies by Glukhovtsev et al.[28,33] on the gas-phase identity S_N2 reactions of halide anions and methyl halides, $X^- + CH_3X$, showed that the calculated G2(+) overall gas-phase barriers for the retention pathway are substantially higher than the corresponding values for back-side attack with inversion of configuration by more than 164.9 kJ mol^{-1}. More recently, Bickelhaupt et al.[33] explored the PESs of the back-side as well as the front-side S_N2 reactions of $X^- + CH_3Y$, with X, Y = F, Cl, Br, and I, using relativistic DFT, and concluded that the front-side S_N2 barriers in all cases were much higher in energy (ca. 160 kJ mol^{-1}), due to a more severe steric repulsion as a result of the proximity between the Nu and the leaving group.

In the present study, the reaction of CH_3CH_2Cl with one representative base, F^-, is used to check the G2(+) energy difference between the back-side and front-side S_N2 TSs with its competitive *anti*- and *syn*-E2 TSs. Figure 1 presents the structures of these two pairs of TSs and their G2(+) energies relative to separated reactants, F^- and CH_3CH_2Cl. The calculated G2(+) energy of the *syn*-E2-TS is higher than that of the *anti*-E2-TS by 45.0 kJ mol^{-1}, signifying that the *anti*-elimination pathway is energetically favorable in the E2 reaction of F^- with CH_3CH_2Cl. The calculated overall G2(+) barrier for the back-side S_N2 reaction is much lower (by 169.2 kJ mol^{-1}) than that of the front-side S_N2 reaction of F^- with CH_3CH_2Cl, indicating that the inv-S_N2 pathway is much more favorable than the retention one. So, we will only focus on the back-side S_N2 and *anti*-E2 TS structures in the following discussion.

The key TS structural parameters for the back-side S_N2 transition states are the distance between the attacking atom and the central carbon, $X\cdots C^\alpha$, and the distance between the central carbon atom and the leaving chloride ion, $C^\alpha\cdots Cl$. These distances can be better assessed by their bond orders $n^\ddagger(X-C^\alpha)$ and $n^\ddagger(C^\alpha-Cl)$ (see Table 2). It is found that the

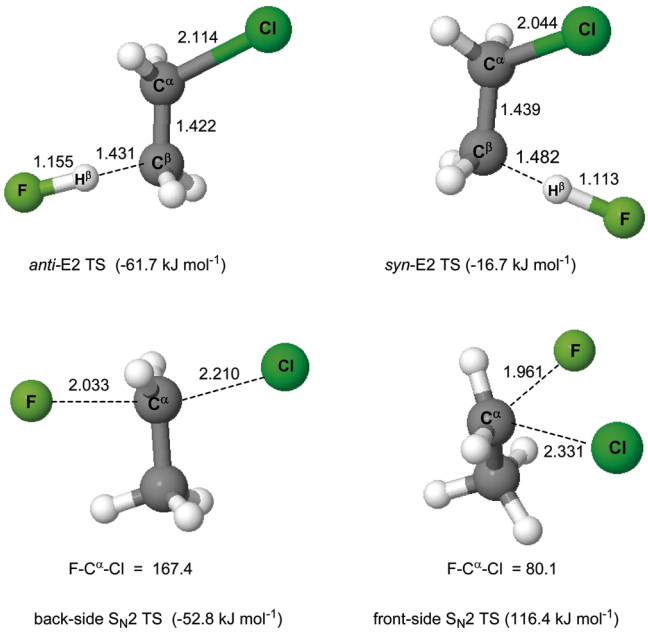


Figure 1. MP2/6-31+G(d) optimized structures for the *anti* and *syn* E2 TSs, back-side and front-side S_N2 TSs of ethyl chloride with fluoride ion, in which the bond lengths and angles are in angstroms and degrees, respectively. The numbers in parentheses are the G2(+) energies relative to the separated reactants, CH_3CH_2Cl and F^-.

Table 2. Selected Geometrical Parameters (in Normal Font) and the Bond Order, $\Delta n^{\ddagger}$(in Bold Font), in the S_N2 TS Structures $[X\cdots Et\cdots Cl]^{-\ddagger}$

X^-	$r(X-C^{\alpha})$	$n^{\ddagger}(X-C^{\alpha})$	$r(C^{\alpha}-Cl)$	$n^{\ddagger}(C^{\alpha}-Cl)$
F^-	2.033	**0.358**	2.210	**0.499**
Cl^-	2.354	**0.392**	2.381	**0.375**
Br^-	2.456	**0.440**	2.404	**0.361**
HO^-	2.202	**0.279**	2.155	**0.547**
HS^-	2.540	**0.304**	2.264	**0.456**
HSe^-	2.610	**0.339**	2.253	**0.465**
NH_2^-	2.384	**0.218**	2.103	**0.597**
PH_2^-	2.745	**0.230**	2.191	**0.515**
AsH_2^-	2.791	**0.251**	2.165	**0.538**
CH_3^-	2.647	**0.155**	2.060	**0.641**
SiH_3^-	2.808	**0.216**	2.190	**0.516**
GeH_3^-	2.730	**0.266**	2.180	**0.525**

S_N2 TS structures have decreasing $n^{\ddagger}(X-C^{\alpha})$ values on going from left to right of a given row in the periodic table and increasing when going down a group, showing that there is an earlier TS for the Nu with stronger basicity. Unexpectedly, the $C^{\alpha}-Cl$ distances for the S_N2 TSs do not increase monotonically from top to bottom for the groups 14−16 Nus with decreasing PA values. For example, the $C^{\alpha}-Cl$ distance in the S_N2 TS $[HS\cdots Et\cdots Cl]^{-\ddagger}$ is slightly longer than that in the $[HSe\cdots Et\cdots Cl]^{-\ddagger}$. This trend is also observed in groups 14 and 15. The magnitude of geometrical deformation of TSs can be described by their deformation energy, ΔH_{def}, defined as the enthalpy change accompanying the transformation from equilibrium reactant structures to the corresponding TS, which is also called "activation strain" by Bickelhaupt. Obviously, higher deformation energy for the S_N2 reaction arises mainly from the more cleaved $C^{\alpha}-Cl$ bond in the TS, that is, the smaller $n^{\ddagger}(C^{\alpha}-Cl)$ value, which is supported by the good linear correlation ($R^2 = 0.992$) for the plot of $\Delta H_{def}(S_N2)$ against $n^{\ddagger}(C^{\alpha}-Cl)$.

For the E2 reactions, the main geometrical character of TSs can be described by $n^{\ddagger}(X-H^{\beta})$, $n^{\ddagger}(H^{\beta}-C^{\beta})$, $n^{\ddagger}(C^{\alpha}-C^{\beta})$, and $n^{\ddagger}(C^{\alpha}-Cl)$ (see Table 3), in which the elongation of $C^{\beta}-H^{\beta}$ and $C^{\alpha}-Cl$ bonds will significantly contribute to the deformation energy of E2 TS, and the smaller sum of $n^{\ddagger}(C^{\beta}-H^{\beta})$ and $n^{\ddagger}(C^{\alpha}-Cl)$ will result in a large $\Delta H_{def}(E2)$ value, leading to a reasonable correlation ($R^2 = 0.953$) for $n^{\ddagger}(C^{\beta}-H^{\beta}) + n^{\ddagger}(C^{\alpha}-Cl)$ against $\Delta H_{def}(E2)$. Data in Table 3 show that there are smaller $n^{\ddagger}(C^{\alpha}-C^{\beta})$ and larger $n^{\ddagger}(C^{\alpha}-Cl)$ values for the first-row bases (X= CH_3, NH_2, HO, and F) with stronger basicity. When the weaker bases, such as HSe^-, Cl^-, and Br^-, attack the H^{β} on the substrate, there are more product-like characteristics, as evidenced by the larger $n^{\ddagger}(C^{\alpha}-C^{\beta})$ and smaller $n^{\ddagger}(C^{\alpha}-Cl)$ values in those E2 TSs. In fact, the reasonable correlations existing for PA versus $n^{\ddagger}(C^{\alpha}-C^{\beta})$ ($R^2 = 0.963$) and versus $n^{\ddagger}(C^{\alpha}-Cl)$ ($R^2 = 0.957$) indicate a more product-like TS for the weaker bases.

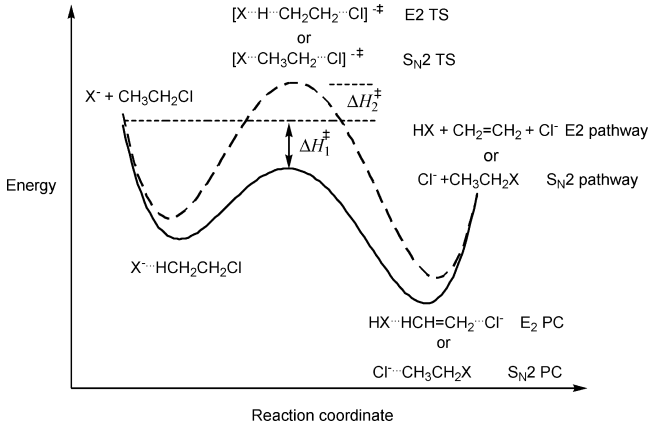
3.3. The Barrier Heights for the S_N2 and E2 Reactions. It is well-known that the PESs for both of the S_N2 and E2 reactions have the shape of a double well, as shown in Scheme 2. The first step in the present study involves the initial exothermic formation of a reactant ion−molecule complex, $X^-\cdots CH_3CH_2Cl$. This complex is predestined to react further via a favorable back-side S_N2 or *anti*-E2 pathway. The reaction then proceeds via the S_N2 or *anti*-E2 TS, yielding the product complex (PC) $Cl^-\cdots CH_3CH_2X$ for the S_N2 pathway or $HX\cdots CH_2{=}CH_2\cdots Cl^-$ for the E2 pathway. These complexes can decompose into the products $Cl^- + CH_3CH_2X$ or $CH_2{=}CH_2 + HX + Cl^-$, in which the leaving group Cl^- and the conjugate acid HX can form a stable complex as the most stable products in the E2 pathway. Nibbering previously pointed out that the overall barrier, $\Delta H^{\ddagger}$, is decisive for the rate of chemical reactions in the gas phase, particularly if they occur under low-pressure conditions in which the reacting system is (in good approximation) thermally isolated.[57,58] This is the reason why we only discuss the overall barriers in the following discussion.

The G2(+) overall barriers for the back-side S_N2 and *anti*-E2 reactions with CH_3CH_2Cl, $\Delta H^{\ddagger}(S_N2)$, and $\Delta H^{\ddagger}(E2)$ are collected in Table 4. Bickelhaupt and his co-worker[30] found that the sequence given by $\Delta H^{\ddagger}$ in the S_N2 reactions of halides with methyl halides follows the decreasing order $I^- > Br^- > Cl^- > F^-$; that is, the reactivity order of halide anion decreases from top to bottom in group 17. This trend is also observed in the present S_N2 and E2 reactions of ethyl chloride; that is, the $\Delta H^{\ddagger}(S_N2)$ and $\Delta H^{\ddagger}(E2)$ values always increase within each group as we go down the periodic table. Our calculation results are in accord with the existing experimental data. For example, Bierbaum et al.[59] measured the rate coefficients for the substitution reactions of a series of anions toward CH_3I by using FA-SIFT techniques and reported the following reactivity order: $F^- > Cl^- > Br^-$ and $HO^- > HS^-$; Anderson et al.[60] investigated the gas PH_2^- reactions with a series of neutral substrates including CH_3Y (Y = Cl, Br, and I) using the

Table 3. Selected Geometrical Parameters (in Normal Font) and the Bond Order, $\Delta n^{\ddagger}$(in Bold Font), in the E2 TS Structures $[X\cdots H^{\beta}\cdots CH_2CH_2\cdots Cl]^{-\ddagger}$

X^-	$r(X-H^{\beta})$	$n^{\ddagger}(X-H^{\beta})$	$r(C^{\beta}-H^{\beta})$	$n^{\ddagger}(C^{\beta}-H^{\beta})$	$r(C^{\alpha}-C^{\beta})$	$n^{\ddagger}(C^{\alpha}-C^{\beta})$	$r(C^{\alpha}-Cl)$	$n^{\ddagger}(C^{\alpha}-Cl)$
F^-	1.155	**0.700**	1.431	**0.572**	1.422	**1.368**	2.114	**0.586**
Cl^-	1.514	**0.678**	1.494	**0.515**	1.378	**1.584**	2.489	**0.313**
Br^-	1.636	**0.717**	1.546	**0.472**	1.369	**1.632**	2.644	**0.242**
HO^-	1.297	**0.581**	1.363	**0.641**	1.453	**1.234**	1.975	**0.738**
HS^-	1.635	**0.612**	1.445	**0.559**	1.396	**1.492**	2.306	**0.425**
HSe^-	1.738	**0.649**	1.463	**0.542**	1.386	**1.542**	2.421	**0.351**
NH_2^-	1.444	**0.491**	1.325	**0.683**	1.46	**1.205**	1.938	**0.785**
PH_2^-	1.795	**0.530**	1.399	**0.604**	1.411	**1.419**	2.204	**0.504**
AsH_2^-	1.864	**0.577**	1.398	**0.605**	1.402	**1.462**	2.291	**0.436**
CH_3^-	1.642	**0.399**	1.298	**0.714**	1.461	**1.201**	1.93	**0.796**
SiH_3^-	1.903	**0.496**	1.399	**0.604**	1.413	**1.410**	2.193	**0.513**
GeH_3^-	1.895	**0.562**	1.411	**0.592**	1.402	**1.462**	2.28	**0.444**

Scheme 2. Schematic Potential Energy Diagrams for the Gas-Phase E2 and S_N2 Reactions, in Which the Plain Line Is for Negative Overall Barrier ($\Delta H_1^{\ddagger}$) and the Dashed Line Is for Positive Overall Barrier ($\Delta H_2^{\ddagger}$)



FA technique. These reactions were compared to those for the reactions of NH_2^-. Many similarities exist between the reactions of phosphide and those of amide, but the former reacts much less efficiently than the latter.

The overall barriers in Table 4 show that all of the S_N2 and E2 TSs for the four first-row bases have energies ranging from 36.3 to 63.5 kJ mol^{-1} below that of the separated reactants. The lower overall barriers are consistent with the higher complexation energies between X^- with higher PA values and CH_3CH_2Cl. For example, the G2(+) complexation enthalpy for the complex formed between F^- and CH_3CH_2Cl, $F^-\cdots CH_3CH_2Cl$, is -75.0 kJ mol^{-1} with respect to the separated reactants, leading to the overall barrier, $\Delta H^{\ddagger}(S_N2) = -52.8$ kJ mol^{-1} and $\Delta H^{\ddagger}(E2) = -61.7$ kJ mol^{-1}, much lower than the previous values of -28.0 and -23.8 kJ mol^{-1}, respectively, reported by Gronert et al.,[16] calculated at the MP4SDQ/6-31+G(d,p)//HF/6-31+G(d) level. A similar situation is also found in S_N2 TS $[H_2P\cdots Et\cdots Cl]^{-\ddagger}$, in which the G2(+) $\Delta H^{\ddagger}(S_N2)$ value is -6.1 kJ mol^{-1}.

The $\Delta H^{\ddagger}$ sequences for the S_N2 reactions show the following decreasing order:

$$HO < F < NH_2 < CH_3 < PH_2 < HS < AsH_2 \approx SiH_3 < GeH_3 < HSe < Cl < Br$$

The above order holds in most cases for the corresponding E2 reactions, but there are some deviations:

$$HO < F < NH_2 < CH_3 < PH_2 < HS < SiH_3 < AsH_2 < Cl < GeH_3 < HSe < Br$$

These two trends appear to be related to the reactivity of these anions toward ethyl chloride, which match the available experimental data.[10,59,60] It is worth noting that the differences of the S_N2 or E2 barriers involving the Nus (or bases) of similar basicity could be considerable. Moreover, the weaker base may show higher reactivity than the stronger one. For example, the PA values of HO^- and CH_3^- are 1631.3 and 1746.6 kJ mol^{-1}, respectively, but the calculated $\Delta H^{\ddagger}(S_N2)$ and $\Delta H^{\ddagger}(E2)$ values are -56.8 and -63.5 kJ mol^{-1} for HO^-, and -43.4 and -36.3 kJ mol^{-1} for CH_3^-, indicating that (1) CH_3^- has much lower reactivity than HO^- in both S_N2 and E2 reactions; (2) with hydroxide anion as the Nu, the elimination is more favorable than the substitution, but the two processes should be competitive; and (3) CH_3^- has preference for the back-side S_N2 pathway. These results imply that there are other factors for determining the reactivity of the Nu in the S_N2 or E2 reactivity in addition to its basicity.

Even though the enhanced reactivity of hydroxide anion could be explained by the hard and soft acids and bases (HSAB) principle, here the idea of electron reorganization is used to rationalize its stronger reactivity in the S_N2 and E2 reactions. In the S_N2 reaction of HO^- with CH_3CH_2Cl, the net charge on the HO moiety decreases from -1.00e to -0.33e (in S_N2 product CH_3CH_2OH). The change of population is much smaller than that in the S_N2 reaction of CH_3^- with CH_3CH_2Cl, in which the electron population on the CH_3 moiety shifts from 10.00e to 8.99e in going from reactant CH_3 to product $CH_3CH_2CH_3$, implying that much more electron on the HO will be retained than that on the CH_3 moiety in the product when S_N2 reaction occurs, which in turn will lead to a much lower S_N2 barrier for the reaction of HO^- with CH_3CH_2Cl.

This rationalization can be extended to the E2 reactions. When HO^- initiates the proton transfer of the *anti*-E2 pathway, the electron density on the HO^- moiety changes only from -1.00e to -0.47e in the product H_2O. In contrast, the E2 reaction of CH_3^- with CH_3CH_2Cl results in much more change of net charge, and the population on the CH_3 moiety shifts from 18.00e to 16.79e, leading to a significant

Table 4. Calculated G2(+) Reaction Barriers Relative to the Separated Reactants, $\Delta H^{\ddagger}$, Deformation Energies, ΔH_{def}, and Actual Interaction, ΔH_{int}, between the Deformed Reactants in the TS, and Total Reaction Enthalpy Changes, ΔH, for the Gas-Phase Reactions $X^- + CH_3CH_2Cl$[a]

X^-	$\Delta H^{\ddagger}(S_N2)$	$\Delta H^{\ddagger}(E2)$	$\Delta H_{def}(S_N2)$	$\Delta H_{def}(E2)$	$\Delta H_{int}(S_N2)$	$\Delta H_{int}(E2)$	$\Delta H(S_N2)$	$\Delta H(E2)$
F^-	−52.8	−61.7	87.8	146.4	−140.6	−208.1	−134.3	−80.8
Cl^-	18.1	61.3	140.2	262.3	−122.1	−201.0	0.0	72.3
Br^-	35.0	93.5	148.0	312.2	−113.0	−218.7	36.0	115.3
HO^-	−56.8	−63.5	68.9	88.1	−125.7	−151.6	−209.2	−161.4
HS^-	2.7	45.6	104.0	201.4	−101.3	−155.8	−78.2	−2.2
HSe^-	17.4	71.9	102.0	237.0	−84.5	−165.1	−40.5	40.7
NH_2^-	−50.3	−50.5	54.0	108.5	−104.3	−158.9	−276.1	−219.0
PH_2^-	−6.1	35.3	80.7	162.0	−86.8	−126.7	−165.5	−69.1
AsH_2^-	5.1	54.6	73.5	182.2	−68.4	−127.6	−124.6	−30.7
CH_3^-	−43.4	−36.3	40.4	57.4	−83.8	−93.7	−356.7	−276.7
SiH_3^-	5.2	52.3	83.4	167.5	−78.2	−115.2	−208.4	−92.9
GeH_3^-	16.8	67.7	80.7	193.6	−63.9	−125.9	−150.7	−47.8

[a] All energies are in kJ mol^{-1}.

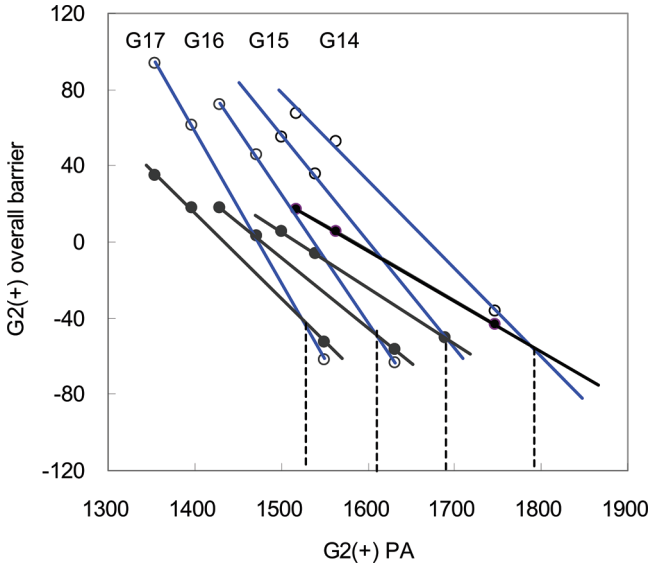


Figure 2. Plot of the G2(+) overall barrier (kJ mol^{-1}) vs the PAs (kJ mol^{-1}) along each column of the peordic table for *anti*-E2 reactions (blue line) and for back-side S_N2 reactions (black line).

electron reorganization and higher E2 barrier in the reaction of CH_3^- with ethyl chloride.

3.4. Correlation of E2 and S_N2 Barrier Height with PA and EN. Figure 2 shows the relation between the overall barrier for E2 reactions, $\Delta H^{\ddagger}(E2)$, and PA for various groups. It can be inferred from this figure that there is no general and straightforward relationship between $\Delta H^{\ddagger}(E2)$ and PA for all of the bases. Instead, there is an excellent linear relationship ($R^2 \approx 1.00$) within each column of the periodic table (see Figure 2). Similar trends occur for the corresponding S_N2 reactions (see also Figure 2).

In the early study of identity proton transfer reaction between simple hydrides ($AH + A^- \rightarrow A^- + AH$), Gronert[61] found a stronger correlation between the EN of central atom in A and the barrier to the proton transfer. He interpreted these results in terms of a model where the TS was dominated by the triple ion valence bond resonance configuration, $[A\cdots H\cdots A]^{-\ddagger} \leftrightarrow [A^-H^+A^-]^{\ddagger}$, where the transferring proton and base carried full charges. Obviously, this resonance form would be more stable when A is highly electronegative. The stabilization of TSs by resonance also seems to be applicable in the present S_N2 and E2 reactions. In the E2 reactions of CH_3CH_2Cl, proton transfer still takes place, and this transfer is accompanied by the leaving of a chloride ion. In the S_N2 reaction, the TS structure can be viewed as a resonance form similar to that in the proton transfer reaction. So there may be some type of relationship between the EN of attacking atom and the overall barriers. Here, the revised EN scales V_x (eq 5), suggested by Luo and Benson,[62] for the 12 attacking atoms covering groups 14−17 of the periodic table are used to correlate with the overall barriers of S_N2 and E2, where n_x is the number of valence electrons, and r_x is the covalent radius from ref 63:

$$V_x = n_x/r_x \quad (5)$$

The stabilization of the TS by more electronegative Nus can also be understood in terms of the bonding and, especially, the nonbonding orbital in the three-center four-electron picture of these species, which has been discussed in detail by Pierrefixe et al.[64−66] The occupied nonbonding MO has high amplitudes on the terminal groups (nucleophile and leaving group) and will be stabilized if these groups become more electronegative.

As in the case of PA, there are also good linear correlations ($R^2 = 0.99-1.00$) between EN and $\Delta H^{\ddagger}(E2)$ or $\Delta H^{\ddagger}(S_N2)$ for every column in the periodic table. Because both PA and EN are important for determining the E2 and S_N2 barriers, we should correlate the overall barrier with both PA and EN. We report here a two-parameter treatment of our results for all E2 (eq 1) or S_N2 reactions (eq 2) by multiple linear regression analysis. The results provide reasonable correlations (see eqs 6 and 7), indicating that it is possible to approximately predict the S_N2 and E2 overall barriers for normal Nu toward ethyl chloride on the basis only of the PA value of Nu and the EN value of the attacking atom. For example, the predicted values by using eqs 6 and 7 for the reaction of CH_3O^- and CH_3CH_2Cl are $\Delta H^{\ddagger}(E2) = -46.9$ kJ mol^{-1}, and $\Delta H^{\ddagger}(S_N2) = -51.4$ kJ mol^{-1}, which are very close to the calculated G2(+) ones, −48.3 and −50.7 kJ mol^{-1}. Using eq 5, the covalent radius of oxygen is 0.73 Å.[63]

$$\Delta H^{\ddagger}(E2) = -0.38PA - 17.69EN + 713.00 \; (R^2 = 0.986, n = 12) \quad (6)$$

$$\Delta H^{\ddagger}(S_N2) = -0.23PA - 8.89EN + 389.54 \; (R^2 = 0.974, n = 12) \quad (7)$$

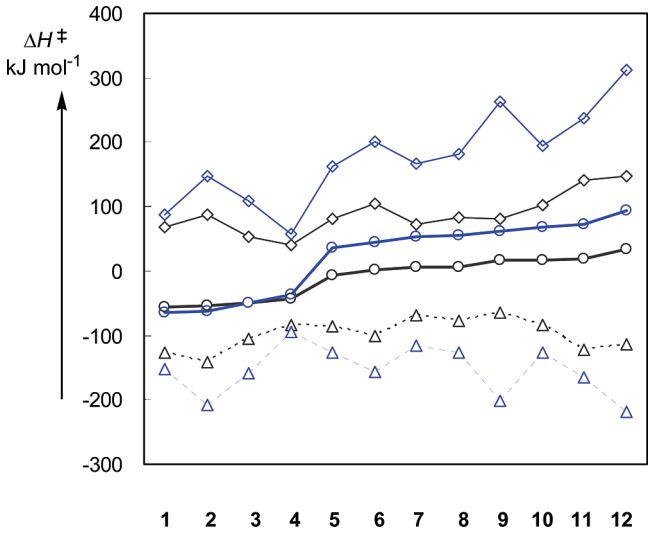


Figure 3. Variations of X^- with deformation energy (plain line), interaction between the deformed reactants (dashed line), and overall barrier (bold line) for the S_N2 (black line, X = 1, HO; 2, F; 3, NH_2; 4, CH_3; 5, PH_2; 6, HS; 7, AsH_2; 8, SiH_3; 9, GeH_3; 10, HSe; 11,Cl; 12, Br) and E2 reactions (blue line, X = 1, HO; 2, F; 3, NH_2; 4, CH_3; 5, PH_2; 6, HS; 7, SiH_3; 8, AsH_2; 9, Cl; 10, GeH_3; 11, HSe; 12, Br) of X^- with EtCl along the increasing overall barrier trend.

3.5. Relationship of E2 and S_N2 Barrier Heights with Deformation Energy of the TS. In our previous papers,[35,36] it was found that there exist reasonable linear correlations between the deformation energies and overall barriers in the S_N2 and E2 reactions. Inspection of the data in Table 4 shows that the correlation only exists for the E2 reactions when the attacking atoms are on the same column in the periodic table, in which the deformation energies increase from top to bottom, whereas there is a linear correlation only for the group 17 Nus (halide anions) in the S_N2 reactions due to the irregular $n^{\ddagger}(C^{\alpha}-Cl)$ values when going down the groups 14−16. These results imply that the deformation energy does not generally dominate the overall barrier for both of the S_N2 and E2 reactions, and other factors need to be considered. Here, we use the idea of energy decomposition introduced by Bickelhaupt[67] to analyze the factors determining the barrier heights of S_N2 and E2 reactions, that is, the activation strain model. In this model, the overall reaction barrier can be partitioned into deformation energy (ΔH_{def}) and interaction (ΔH_{int}) between the deformed reactants in the TS. Figure 3 illustrates the variations of ΔH_{def} (normal line), ΔH_{int} (dashed line), and their sum, $\Delta H^{\ddagger}$ (bold line), for a series of E2 reactions (eq 1, blue line) and S_N2 reactions (eq 2, black line) along the increasing $\Delta H^{\ddagger}$ values. Generally speaking, a larger deformation energy will destabilize the TS and raise the overall barrier, but when the interaction between the deformed reactants is stronger, the barrier order could be reversed.

For the S_N2 reactions of the groups 14−16 Nus, the deformation energies from the second-row elements are slightly higher than those from the third-row ones by up to 7.2 kJ mol^{-1} induced by the longer $C^{\alpha}-Cl$ bond in the former cases. The overall barriers, $\Delta H^{\ddagger}(S_N2)$, are still increasing from top to bottom due to the stronger interaction between deformed reactants, which may be explained by the much higher ECA value of second-row Nus. Moreover, the interactions, ΔH_{int}, are found to correlate well with the ECAs in each column.

For all of the present E2 reactions, the deformation energies are in general much larger than the corresponding S_N2 reactions due to the fact that two bond cleavages are involved in the E2 TS, which has been also analyzed and pointed out by Bickelhaupt.[67] Meanwhile, the interactions between the deformed reactants in the E2 TS are also stronger than those in the S_N2 reactions except CH_3^- because of its larger PA value than the corresponding ECA one of X^-. For the CH_3^-case, the smaller ΔH_{int} value of −93.7 kJ mol^{-1} for the E2 TS, $[H_3C\cdots H^{\beta}\cdots CH_2CH_2\cdots Cl]^{-\ddagger}$, can be rationalized by the much smaller $n^{\ddagger}(C-H^{\beta})$ value, leading to weaker interaction between $CH_3^-$and H^{β} in the E2 TS.

3.6. Correlation of E2 and S_N2 Barrier Heights with Reaction Enthalpy. As shown in previous work,[68] the exothermicity of the reaction of nucleophile with a single substrate reflects the thermodynamic affinity of the nucleophile. Following this idea, the exothermicity trend, in this work, is given by the sequences of the overall enthalpy change, denoted as ΔH_{ovr}, as a function of Nu or base: the more negative is ΔH_{ovr}, the stronger is the exothermicity of the reaction. It can be seen from Table 4 that, for both the S_N2 and E2 reactions with substrate CH_3CH_2Cl, the exothermicity will decrease in going down a group, and the relationship between $\Delta H^{\ddagger}$ and ΔH_{ovr} follows the same trend as PA; that is, the consistency of the kinetics and thermodynamics for the present S_N2 and E2 reactions only exists within each column of the periodic table.

3.7. Competition between S_N2 and E2 Reactions. Comparison of the E2 and S_N2 overall barriers in Table 4 shows that the S_N2 pathway is much more favorable for all of the second- and third-row Nus, which is consistent with previous results. For the first-row bases, E2 dominates for reactions of $F^- + CH_3CH_2Cl$ and $HO^- + CH_3CH_2Cl$, and S_N2 and E2 pathways are competitive for the reaction of $NH_2^- + CH_3CH_2Cl$, whereas S_N2 reaction dominates for $CH_3^- + CH_3CH_2Cl$.

If we combine the correlations between S_N2 and E2 overall barriers and PA values along each column of periodical table, a very clear picture emerges from the analysis of the crossing points. Figure 2 shows that the crossing points from the two series of correlations of $\Delta H^{\ddagger}$ versus PA of X will shift to the right from group 17 to group 14, which implies that the favorable pathway is related to the position of the attacking atom in the periodic table.

4. Conclusions

This work systematically studies the reactions of ethyl chloride with a series of Nus covering the groups 14−17 elements using the G2(+) method. Two competitive reaction pathways, back-side S_N2 and *anti*-E2, are investigated, leading to the following conclusions.

(1) For both the S_N2 and the E2 reactions, the good correlation between G2(+) PAs and overall barriers, $\Delta H^{\ddagger}$, only exists when the attacking atoms belong to the same group in the periodic table. This modifies the previous claim

that $\Delta H^{\ddagger}$ values for the S_N2 and E2 reactions are basically controlled by the PA of bases. Thus, it is more reasonable to make a reference line using the nucleophiles or bases with the attacking atom in the same group when discussing the α-effect in the E2 and S_N2 reactions.

(2) A strong correlation is found between the EN of the attacking atom and the barrier heights of S_N2 and E2 reactions. A higher EN value of X will stabilize the S_N2 and E2 TS by less electron reorganization. Good linear correlation exists for $\Delta H^{\ddagger}$ versus EN within the same column of the periodic table.

(3) Two-parameter equations are derived to connect the S_N2 or E2 overall barriers with the combination of PA and EN values of the attacking atom by multiple linear regression analysis, indicating the importance of both PA and EN in determining the S_N2 or E2 reactivity. Thus, the PA and EN values may be used to predict the overall barrier of the S_N2 or E2 reactions involving normal Nu.

(4) It is found that the good correlation of $\Delta H^{\ddagger}$ versus ΔH_{def} only exists in E2 reactions with attacking atoms in the same group, which deviates from the previous conclusion that there is a general linear relationship between the overall barrier and all Nus or bases.

Acknowledgment. This work is supported by a Strategic Grant (Project No. 7002334) from the City University of Hong Kong.

Supporting Information Available: Cartesian coordinates of all species reported. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lowry, T. H.; Richardson, K. S. *Mechanism and Theory in Organic Chemistry*, 3rd ed.; Harper and Row: New York, 1987.

(2) Jones, M. E.; Ellison, G. B. *J. Am. Chem. Soc.* **1989**, *111*, 1645.

(3) Dhar, M. L.; Hughes, E. D.; Ingold, C. K.; Masterman, S. *J. Chem. Soc.* **1948**, *48*, 2055.

(4) Lum, R. C.; Grabowski, J. J. *J. Am. Chem. Soc.* **1992**, *114*, 9663.

(5) Flores, A. E.; Gronert, S. *J. Am. Chem. Soc.* **1999**, *121*, 2627.

(6) Gronert, S.; Fagin, A. E.; Okamoto, K.; Mogali, S.; Pratt, L. M. *J. Am. Chem. Soc.* **2004**, *126*, 12977.

(7) Gronert, S.; Fagin, A. E.; Wong, L. *J. Am. Chem. Soc.* **2007**, *17*, 5331.

(8) DePuy, C. H.; Bierbaum, V. M. *J. Am. Chem. Soc.* **1981**, *103*, 5034.

(9) DePuy, C. H.; Beedle, E. C.; Bierbaum, V. M. *J. Am. Chem. Soc.* **1982**, *104*, 6483.

(10) DePuy, C. H.; Gronert, S.; Mullin, A.; Bierbaum, V. M. *J. Am. Chem. Soc.* **1990**, *112*, 8650.

(11) Gronert, S.; DePuy, C. H.; Bierbaum, V. M.; DePuy, C. H. *J. Am. Chem. Soc.* **1991**, *113*, 4009.

(12) Ridge, D. P.; Beauchamp, J. L. *J. Am. Chem. Soc.* **1974**, *96*, 637.

(13) Minato, T.; Yamabe, S. *J. Am. Chem. Soc.* **1980**, *107*, 4621.

(14) Bento, A. P.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Theory Comput.* **2008**, *4*, 929.

(15) Gronert, S.; Merrill, G. N.; Kass, S. R. *J. Org. Chem.* **1995**, *60*, 488.

(16) Gronert, S. *J. Am. Chem. Soc.* **1991**, *113*, 6041.

(17) Gronert, S. *J. Am. Chem. Soc.* **1993**, *115*, 652.

(18) Gronert, S.; Kass, S. R. *J. Org. Chem.* **1997**, *62*, 7991.

(19) Gronert, S. *J. Org. Chem.* **1994**, *59*, 7046.

(20) Gronert, S.; Freed, P. *J. Org. Chem.* **1996**, *61*, 9430.

(21) Hu, W.-P.; Truhlar, D. G. *J. Am. Chem. Soc.* **1996**, *118*, 860.

(22) Villano, S. M.; Kato, S.; Bierbaum, V. M. *J. Am. Chem. Soc.* **2006**, *128*, 736.

(23) Su, T.; Chesnavich, W. J. *J. Chem. Phys.* **1982**, *76*, 5183.

(24) Pabis, A.; Paluch, P.; Szala, J.; Paneth, P. *J. Chem. Theory Comput.* **2009**, *5*, 33.

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

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

(27) Miertus, S.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1981**, *55*, 117.

(28) Glukhovtsev, M. N.; Pross, A.; Radom, L. *J. Am. Chem. Soc.* **1995**, *117*, 2024.

(29) Glukhovtsev, M. N.; Pross, A.; Radom, L. *J. Am. Chem. Soc.* **1995**, *117*, 9012.

(30) Bento, A. P.; Bickelhaupt, F. M. *J. Org. Chem.* **2008**, *73*, 7290.

(31) Lee, I.; Kim, C. K.; Lee, B. S. *J. Phys. Org. Chem.* **1995**, *8*, 473.

(32) Uggerud, E. *Chem.-Eur. J.* **2006**, *12*, 1127.

(33) Glukhovtsev, M. N.; Pross, A.; Radom, L. *J. Am. Chem. Soc.* **1996**, *118*, 6273.

(34) Ren, Y.; Yamataka, H. *Org. Lett.* **2006**, *8*, 119.

(35) Ren, Y.; Yamataka, H. *J. Org. Chem.* **2007**, *72*, 5660.

(36) Ren, Y.; Yamataka, H. *J. Comput. Chem.* **2009**, *30*, 358.

(37) Bento, A. P.; Sola, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2005**, *26*, 1497.

(38) Mourik, T. v. *Chem. Phys. Lett.* **2005**, *414*, 364.

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

(40) Houk, K. N.; Gustafson, S. M.; Black, K. A. *J. Am. Chem. Soc.* **1992**, *114*, 8565.

(41) Lee, J. K.; Kim, C. K.; Lee, B. S.; Lee, I. *J. Phys. Chem. A* **1997**, *101*, 2893.

(42) Kim, C. K.; Hyun, K. H.; Kim, C. K.; Lee, I. *J. Am. Chem. Soc.* **2000**, *122*, 2294.

(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gromperts, R.; Strat-

mann, 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.; Cheng, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.01; Gaussian, Inc.: Pittsburgh, PA, 2003; Wallingford, CT, 2004.

(44) NIST Standard reference Database Number 69; http:/webbook.nist.gov/chemistry (accessed Jan 22, 2009).

(45) Decouzon, M.; Gal, J. F.; Gayraud, J.; Maria, P. C.; Vaglio, G. A.; Volpe, P. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 54.

(46) Ervin, K. E.; Lineberger, C. W. *J. Chem. Phys.* **2005**, *122*, 194303.

(47) Mayer, P. M.; Gal, J. F.; Radom, L. *Int. J. Mass Spectrom. Ion Processes* **1997**, *167*−*168*, 689.

(48) Bartmess, J. E.; Hinde, R. J. *Can. J. Chem.* **2005**, *83*, 2005.

(49) Lischka, H.; Köhler, H. J. *J. Am. Chem. Soc.* **1978**, *100*, 5297.

(50) Raghavachari, K.; Whiteside, R. A.; Pople, J. A.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1981**, *103*, 5649.

(51) Ruscic, B.; Berkowitz, J.; Curtiss, L. A.; Pople, J. A. *J. Chem. Phys.* **1989**, *91*, 114.

(52) Klopper, W.; Kutzelnigg, W. *J. Phys. Chem.* **1990**, *94*, 5625.

(53) Carneiro, J. W.; de, M.; Schleyer, P. v. R.; Saunders, M.; Remington, R.; Schaefer, H. F., III; Rauk, A.; Sorensen, T. S. *J. Am. Chem. Soc.* **1994**, *116*, 3483.

(54) Perera, S. A.; Bartlett, J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1995**, *117*, 8476.

(55) Rio, E. Del.; Lopez, R.; Sordo, T. L. *J. Phys. Chem. A* **1998**, *102*, 6831.

(56) Andrei, H. S.; Solcà, N.; Dopfer, O. *Angew. Chem., Int. Ed.* **2008**, *47*, 395.

(57) Nibbering, N. M. M. *Acc. Chem. Res.* **1990**, *23*, 279.

(58) Nibbering, N. M. M. *Adv. Phys. Org. Chem.* **1988**, *24*, 1.

(59) Bierbaurn, V. M.; Grabowski, J. J.; DePuy, C. H. *J. Phys. Chem.* **1984**, *88*, 1389.

(60) Anderson, D. R.; Bierbaum, V. M.; DePuy, C. H. *J. Am. Chem. Soc.* **1983**, *105*, 4244.

(61) Gronert, S. *J. Am. Chem. Soc.* **1993**, *115*, 10258.

(62) Luo, Y. R.; Benson, S. W. *J. Am. Chem. Soc.* **1989**, *111*, 2480.

(63) Winter, M. J. The periodic table on the WWW; http://www.webelements.com, Copyright 1993−2009.

(64) Pierrefixe, S. C. A. H.; Guerra, C. F.; Bickelhaupt, F. M. *Chem.-Eur. J.* **2008**, *14*, 819.

(65) Pierrefixe, S. C. A. H.; Poater, J.; Im, C.; Bickelhaupt, F. M. *Chem.-Eur. J.* **2008**, *14*, 6901.

(66) Pierrefixe, S. C. A. H.; Bickelhaupt, F. M. *Struct. Chem.* **2007**, *18*, 813.

(67) Bickelhaupt, F. M. *J. Comput. Chem.* **1999**, *20*, 114.

(68) Olmstead, W. N.; Brauman, J. I. *J. Am. Chem. Soc.* **1977**, *99*, 4219.

CT900041Y

JCTC Journal of Chemical Theory and Computation

Theoretical Mechanistic Study of the Oxidative Degradation of Benzene in the Troposphere: Reaction of Benzene−HO Radical Adduct with O_2

Santiago Olivella,*,† Albert Solé,‡ and Josep M. Bofill§

Institut de Química Avançada de Catalunya, CSIC, Jordi Girona 18-26, 08034-Barcelona, Catalonia, Spain, and Departament de Química Física, Departament de Química Orgànica, and Institut de Química Teòrica i Computacional, Universitat de Barcelona, Martí i Franquès 1, 08028-Barcelona, Catalonia, Spain

Received February 16, 2009

Abstract: Competing pathways arising from the reaction of hydroxycyclohexadienyl radical (**1**) with O_2, a key reaction in the oxidative degradation of benzene under tropospheric conditions, have been investigated by means of density functional theory (UB3LYP) and quantum-mechanical (UCCSD(T) and RCCSD(T)) electronic structure calculations. The energetic, structural, and vibrational results furnished by these calculations were subsequently used to perform conventional transition-state computations to predict the rate coefficients and evaluate the product yields. The trans stereoisomer of the peroxyl radical (**4**) produced by the O_2 addition to position 2 of benzene ring in radical **1** is energetically more stable than the cis one, although the rate coefficients at 298 K for the formation of both isomers are predicted to be similar. The cyclization of the cis isomer of **4** to a bicyclic allyl radical (**5**) involves calculated barrier heights ($\Delta U^{\ddagger}$, $\Delta E^{\ddagger}$, $\Delta H^{\ddagger}$, and $\Delta G^{\ddagger}$) significantly lower than those of the cyclization of the trans isomer of **4**. This implies that the formation of the cis isomer of **4** can lead to irreversible loss of radical **1** and that the observed chemical equilibrium **1** + O_2 ↔ **4** essentially involves the trans isomer of **4**. Although the reaction enthalpies computed for the O_2 addition to position 4 of benzene ring in radical **1**, affording the cis and trans stereoisomers of a peroxyl radical (**6**), are similar to those for the addition to position 2, the latter addition mode is clearly preferred because it involves lower barrier heights. The barrier heights computed for the cyclization of either the cis or the trans isomers of **6** to a bicyclic radical bearing a peroxy bridge (**7**) are about twice those computed for the cyclization of either the cis or the trans isomers of **4**. Thus, under tropospheric conditions, it is unlikely that the O_2 addition to position 4 of the benzene ring in radical **1** can contribute to the formation of benzene oxidation products.

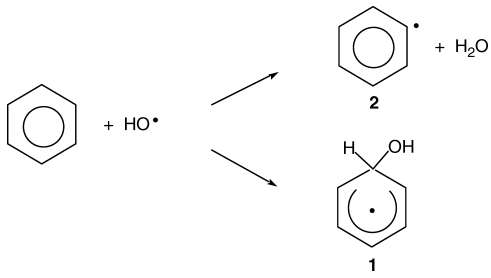
1. Introduction

Benzene is the simplest aromatic hydrocarbon that contributes significantly to the pollution of the troposphere, especially in urban areas of industrialized countries.[1] It is mainly released into the troposphere as a result of anthropogenic activities, such as emissions from burning oil and coal, motor vehicle exhaust, and evaporation of solvents and from gasoline.[2,3] It is now recognized that benzene oxidation reactions may be responsible for a significant fraction of photochemically produced tropospheric ozone.[4] Also, the likely formation of secondary organic aerosols from the oxidation of aromatic hydrocarbons is of considerable concern in connection with human health and the climate.[5]

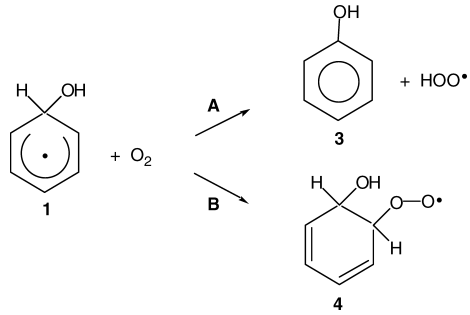
* Corresponding author e-mail: sonqtc@cid.csic.es.
† Institut de Quimica Avançada de Catalunya.
‡ Departament de Química Física i Institut de Química Teòrica i Computacional.
§ Departament de Química Orgànica i Institut de Química Teòrica i Computacional.

10.1021/ct900082g CCC: $40.75
Published on Web 05/21/2009

Scheme 1



Scheme 2

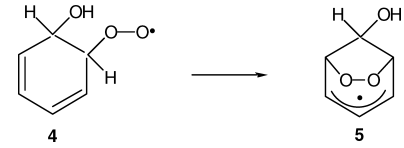


Despite its importance, the knowledge about the tropospheric degradation mechanism of benzene is still scant. Generally, the degradation of benzene in the troposphere is primarily initiated by the addition of hydroxyl radical (HO•) to the aromatic ring, yielding a benzene−HO• adduct: the hydroxycyclohexadienyl radical (**1** in Scheme 1).[6−13] The H-atom abstraction from the aromatic ring leading to formation of phenyl radical (**2** in Scheme 1) is a minor process under tropospheric conditions.[13−15]

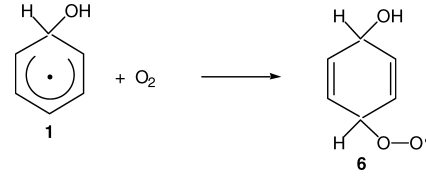
Although the benzene−HO radical adduct **1** initially formed has been found to react more rapidly with NO_2 than with NO, and even more slowly with O_2,[12,16] on the basis of the relative abundance,[17] the latter reaction is the major transformation of this radical in the troposphere.[6−8,18] The presently accepted first elementary steps of this reaction are given in Scheme 2. The reaction can proceed either by abstraction of the H-atom *gem* to HO in **1**, forming phenol (**3**) and HOO• (pathway A), or by O_2 addition to the benzene ring in **1** producing hydroxyl-2,4-cyclohexadienyl-6-peroxyl radical **4** (pathway B). Because of the relatively weak chemical bonding between the C_6H_6OH and OO fragments in peroxyl radical **4**,[6] pathway B is reversible under tropospheric conditions and leads to chemical equilibrium between **1** + O_2 and **4**.[18−22]

The peroxyl radical **4** can undergo various reaction pathways, in addition to the decomposition back to the **1** + O_2 reactants. On the basis of previous theoretical calculations,[23] the ring closure in **4** affording a bicyclic radical (**5** in Scheme 3) appears to be the only plausible unimolecular reaction for peroxyl radical **4**. It is worth noting that the bicyclic radical **5** bears a peroxy bridge and its structure is defined by two fused five- and seven-membered rings, the latter containing a delocalized allyl radical. The subsequent reaction of the bicyclic radical **5** is thought to lead to aromatic ring cleavage forming the principal benzene oxidation products (i.e., glyoxal and butenedial).[24,25]

Scheme 3



Scheme 4



A number of theoretical investigations have focused on characterizing the **1** + O_2 potential energy surface (PES) to explain the experimentally observed branching ratios, thermochemical properties, and rate coefficients.[6,20,22,23,26,27] One of the most thorough theoretical investigations on the primary steps of the benzene oxidation has been published by Lesclaux and co-workers.[22] By using a combination of density functional theory (DFT) and ab initio quantum mechanical calculations with a quadratic correlation (of the Marcus type[28]) between the activation barriers and the reaction enthalpies,[29] Lesclaux and co-workers predicted for the phenol channel (pathway A in Scheme 2) a formation yield of ∼55% in reasonable agreement with the experimental values (25−61%).[1,30−32] Additionally, these authors found that the chemical equilibrium between **1** + O_2 and **4** must essentially involve the trans stereoisomer of **4** (designated by **4**-*trans*), which is less energetic and is formed more rapidly (a factor of about 50) than the cis one (designated by **4**-*cis*). In contrast, the cyclization of **4**-*trans* was calculated to be too slow, as compared to the global rate of irreversible loss of **1** and **4**, whereas it is very fast in the case of **4**-*cis* and can lead readily to benzene oxidation products. However, it must be pointed out that the calculated rate coefficient for the **4**-*cis* formation is a factor of about 10 too low for being consistent with a reasonable yield of oxidation products formed through this reaction channel.[22] Therefore, the possibility of finding another (faster) reaction pathway for the formation of **4**-*cis* deserves a further investigation. Moreover, all previous theoretical studies on the reaction of O_2 with radical **1** leading to peroxyl radicals have focused on the O_2 addition to position 2 of the benzene ring in **1**.[6,20,22,23,26,27] The O_2 addition to position 4 of the benzene ring in **1** affording the hydroxyl-2,5-cyclohexadienyl-4-peroxyl radical (**6** in Scheme 4) appears to be an alternative route[33] that merits a study to elucidate whether or not it plays any relevant role in the tropospheric degradation mechanism of benzene.

New theoretical calculations, using DFT and high level ab initio methods, have been performed in this work aiming to clarify the relative rate coefficients for the formation of the cis and trans isomers of radical **4**, assess the feasibility of the reaction channel leading to the cis and trans isomers of radical **6**, and provide new data on the thermochemistry and kinetics of the reactions of radicals **1**, **4**, and **6**.

2. Computational Details

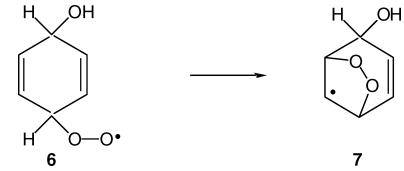
2.1. Electronic Structure Calculations. The geometries of the relevant stationary points (minima and first-order saddle points) on the lowest energy PES of the radical **1** + O_2 reaction system were optimized using analytical gradient procedures,[34] employing DFT calculations. The spin-unrestricted version of the Becke three-parameter hybrid functional[35] combined with the Lee, Yang, and Parr correlation functional,[36] denoted as UB3LYP,[37] was employed with the split-valence 6-31G(d) basis set.[38] All of the stationary points were characterized by their harmonic vibrational frequencies as minima or saddle points. Connections of the transition-state structures between designated minima were confirmed in each case by intrinsic reaction coordinate (IRC)[39] calculations using the second-order algorithm of Gonzalez and Schlegel.[40]

We investigated the effect of adding diffuse sp functions[41] on heavy atoms to the 6-31G(d) basis set on the optimized geometries of the stationary points located at the UB3LYP/6-31G(d) level. The geometries of the reactants (**1** and O_2), five transition structures, and four reaction products involved in reaction pathways shown in Schemes 2 and 3 were obtained with the 6-31G(d) and 6-31+G(d) basis sets and resulted not to differ significantly. Next, the geometries optimized with both basis sets for the reactants and the five transition structures were used in single-point energy calculations at ab initio high levels of theory (see below). The relative energy differences were found to be at variance by 0.5 kcal/mol at most. Furthermore, because the **1** + O_2 reaction system involves two unpaired electrons in O_2 interacting with five delocalized electrons in the π system of radical **1**, we also investigated the effect on the optimized geometries of the transition structures located with the UB3LYP functional of the multireference character expected for these structures. To this end, the geometries of the reactants (**1** and O_2) and five transition structures involved in reaction pathways shown in Schemes 2 and 4 were reoptimized by use of a multiconfiguration self-consistent field wave function of the complete active space (CASSCF) class[42] with the 6-31G(d) basis set. The CAS consisted of 13 electrons and 11 orbitals: the five electrons and five orbitals involved in the π system of the radical **1** unit plus eight electrons and six orbitals of the O_2 unit. The CASSCF-optimized geometries were found to be consistent with those obtained using the B3LYP functional. Details on the two tests are given in the Supporting Information (see Table S1 and Figures S1 and S2). Therefore, the procedure of using the UB3LYP/6-31G(d)-optimized structures in the final single-point calculations at ab initio high levels of theory was deemed safe and adopted throughout this study.

It is well-known that the energy barriers affect the calculated rate coefficients exponentially. Hence, it is crucial to compute accurately the energies of the transition-state structures relative to those of the reactants. Because it is notorious that the UB3LYP functional underestimates the energy barriers calculated for some radical reactions (e.g., in the case of loose transition states and H-atom abstraction reactions),[43] we carried out single-point (frozen core) coupled-cluster[44] calculations including all single and double excitations, based on a reference unrestricted Hartree−Fock (UHF) single determinant, together with a perturbative treatment of all connected triple excitations,[45] designated by UCCSD(T), with the 6-311+G(2df,2p) basis set[46] using the geometries optimized at the UB3LYP/6-31G(d) level. A special difficulty is encountered in the case of the transition-state structures located for the competing reactions shown in Schemes 2−5 because we found a significant difference in the degree of spin contamination shown by the UHF wave function underlying the UCCSD(T) calculations. In fact, the expected values of the spin-squared operator for the UHF/6-311+G(2df,2p) wave function (designated by $\langle S^2 \rangle$) of the transition-state structures calculated for these reactions were ranging between 1.14 and 2.19 (see Table S3, Supporting Information). Therefore, all of the energies were refined by performing single-point energy calculations on the UB3LYP geometries using (frozen core) partially spin-adapted CCSD(T) calculations based on a restricted open-shell Hartree−Fock (ROHF) reference single determinant,[47] designated by RCCSD(T), to avoid the spin contamination problem of the UCCSD(T) calculations.[48]

As noted above, some of the transition states involved in the **1** + O_2 reaction system may have appreciable multireference character and, therefore, may not be well treated with a single-reference based method such as RCCSD(T). To test this suspicion, we computed the T_1 diagnostic values at the RCCSD/6-311+G(2df,2p) level, based on the open-shell formalism of Jayatilaka and Lee,[49] for all of the open-shell species considered in this study (see Table S3, Supporting Information). The T_1 diagnostic gives a qualitative assessment of the significance of nondynamical electron correlation: the larger is the T_1 diagnostic value, the less reliable are the results of the single-reference coupled cluster wave function. For example, the RCCSD method is considered somewhat less reliable if the T_1 diagnostic value is larger than 0.044.[43,50] Examining Table S3 (Supporting Information), we see that all species have T_1 diagnostic values ranging between 0.015 and 0.040 except the transition structure **TS1′**. Thus, our computed RCCSD(T) energy results for **TS1′** may not be entirely reliable, although surely not unreasonable. Fortunately, the energy of **TS1′** is of lesser importance to this study. It is clear then that for all species except **TS1′** our RCCSD(T) results should be reasonably reliable. To provide additional support to this assertion, single-point second-order multiconfigurational perturbation theory calculations (CASPT2),[51] based on the CASSCF(13,11) reference function, were carried out with the 6-31G(d) basis set for the reactants (**1** and O_2) and five transition structures relevant to reactions shown in Schemes 2 and 4. The CASSCF(13,11)/6-31G(d)-optimized geometries were used in these CASPT2 calculations. As shown in Table S2 (Supporting Information),

Scheme 5



6 **7**

the relative energy orderings of these transition structures determined from the CASPT2/6-31G(d) and RCCSD(T)/6-311+G(2df,2p) calculations compare reasonably well.

Zero-point vibrational energies (ZPVEs) were determined from unscaled harmonic vibrational frequencies. Thermal corrections to enthalpy and Gibbs energy values were obtained assuming ideal gas behavior from the unscaled harmonic frequencies and moments of inertia by conventional methods.[52] A standard pressure of 1 atm was taken in the absolute entropy calculations.

All of the UB3LYP and UCCSD(T) calculations were carried out by using the Gaussian 03 program package,[53] whereas the MOLPRO 98 program package[54] was employed for the RCCSD(T) and T_1 diagnostic computations. The CASSCF and CASPT2 calculations were performed by using the GAMESS[55] and MOLCAS-6[56] program packages, respectively.

2.2. Rate Coefficient and Equilibrium Constant Calculations. It is well-known that the theoretical rate coefficient of a reaction is extremely sensitive to the value of the reaction energy barrier. For instance, a change of only 1.4 kcal/mol on the calculated energy barrier causes a change of about a factor of 10 on the calculated rate coefficient.[22] With the main purpose of ascertaining the reliability of the energy barriers obtained from both the UCCSD(T) and the RCCSD(T) calculations, the rate coefficient, k, of the competing reactions shown in Schemes 2−5 was evaluated by using the conventional transition-state theory equation:[57]

$$k = \Gamma \frac{k_b T}{h} \frac{Q_{TS}}{Q_R} e^{-(E_{TS}-E_R)/RT} \quad (1)$$

where Q_{TS} is the partition function of the transition state; Q_R is the product of the partition functions of the reactants; E_{TS} and E_R are the total energy plus the ZPVE of the transition state and reactants, respectively; k_b is the Boltzmann constant; R is the ideal gas constant; T is the absolute temperature; and Γ is the tunneling factor.

According to the standard formulas,[52] the Q's were evaluated using the UB3LYP/6-31(d) geometries and harmonic vibrational frequencies, while the E's were taken as the ZPVE-corrected UCCSD(T)/6-311+G(2df,2p) and RCCSD(T)/6-311+G(2df,2p) energies. The Γ's were evaluated by zero-order approximation to the vibrationally adiabatic PES model with zero curvature.[58] In this approximation, the tunneling is assumed to occur along a unidimensional minimum energy path. The potential energy curve is approximated by an unsymmetrical Eckart potential energy barrier[59] that is required to go through the ZPVE corrected energy (denoted as E) of the reactants, transition state, and products. The equations that describe the Eckart potential energy function were adapted from Truong and Truhlar.[58] Solving the Schroedinger equation for the Eckart function yields the transmission probability, $\kappa(E)$. Γ is then obtained by integrating the respective $\kappa(E)$ over all possible energies:

$$\Gamma(T) = \frac{1}{k_b T} e^{(E_{TS}-E_R)/k_b T} \int_0^\infty e^{-E/k_b T} \kappa(E)\, dE \quad (2)$$

For the reactions of radical **1** with O_2 leading to the formation of peroxyl radicals **4** and **6**, the equilibrium constants expressed in concentration units (denoted as K_c) were evaluated by using the standard formulas:[60]

$$K_c = K_p R' T \quad (3)$$

$$RT \ln K_p = -\Delta G_T^0 \quad (4)$$

where R' is the ideal gas constant in liter atmosphere units, that is, 0.082 L atm/(mol·K), K_p is the equilibrium constant expressed in pressure units, and ΔG_T^0 is the standard Gibbs energy change at 1 atm.

3. Results and Discussion

Selected geometrical parameters of the most relevant structures concerning the stationary points located on the ground-state PES of the **1** + O_2 reaction system at the UB3LYP/6-31G(d) level are shown in Figures 1−6. The Cartesian coordinates of all structures reported in this Article are available as Supporting Information. Total energies computed at UB3LYP, UCCSD(T), and RCCSD(T) levels of theory using the UB3LYP/6-31G(d)-optimized geometries, as well as the ZPVEs, thermal corrections to enthalpy, and Gibbs energy, for all structures are collected in Table S4 (Supporting Information). Tables 1−5 give the relative energies (ΔU), calculated at the UB3LYP, RCCSD(T), and UCCSD(T) levels, the relative energies at 0 K (ΔE(0 K)), and the relative enthalpies (ΔH(298 K)) and Gibbs energies (ΔG(298 K)) at 298 K, calculated at the RCCSD(T) and UCCSD(T) levels, for the stationary points involved in each reaction pathway considered in the present study. Figures 7 and 8 display schematic Gibbs energy profiles of the relevant reaction pathways concerning the O_2 addition to positions 2 and 4 of the benzene ring in radical **1** and the subsequent ring closure of the peroxyl radicals formed. Finally, the values of Γ and k at 298 K for the bimolecular and unimolecular reactions are summarized in Tables 6 and 7, respectively.

3.1. H-Atom Abstraction by O_2 in Hydroxycyclohexadienyl Radical Affording Phenol. Table 1 gives the values of ΔU, ΔE(0 K), ΔH(298 K), and ΔG(298 K) calculated at different levels of theory for the relevant stationary points for the reaction pathway **A** in Scheme 2. In agreement with earlier UB3LYP/6-31(+)G(d) calculations by Ghigo and Tonachini,[26] we found two transition structures for this reaction channel (labeled as **TS1** and **TS1′** in Figure 1). Their geometries differ one from the other essentially in the orientation of the O−O and O−H bonds relative to the benzene cycle. Furthermore, **TS1′** shows an intermolecular hydrogen bond between an oxygen atom of the O_2 unit and the hydrogen atom of the OH group. However, the UB3LYP/6-31G(d) calculations predict that the total energy of **TS1′** is 1.4 kcal/mol higher than that of **TS1** (see Table 1). To investigate the origin of this unexpected result, we performed an analysis of the electron density in **TS1** and **TS1′** within the framework of the topological theory of an atoms in molecules (AIM).[61] The AIM topological analysis of the electron density in **TS1′** revealed the presence of a bond critical point between one of the two oxygen atoms of the O_2 unit and the hydrogen atom of the OH group, with an electron density of 0.0289 e, which can be associated with the aforementioned intermolecular

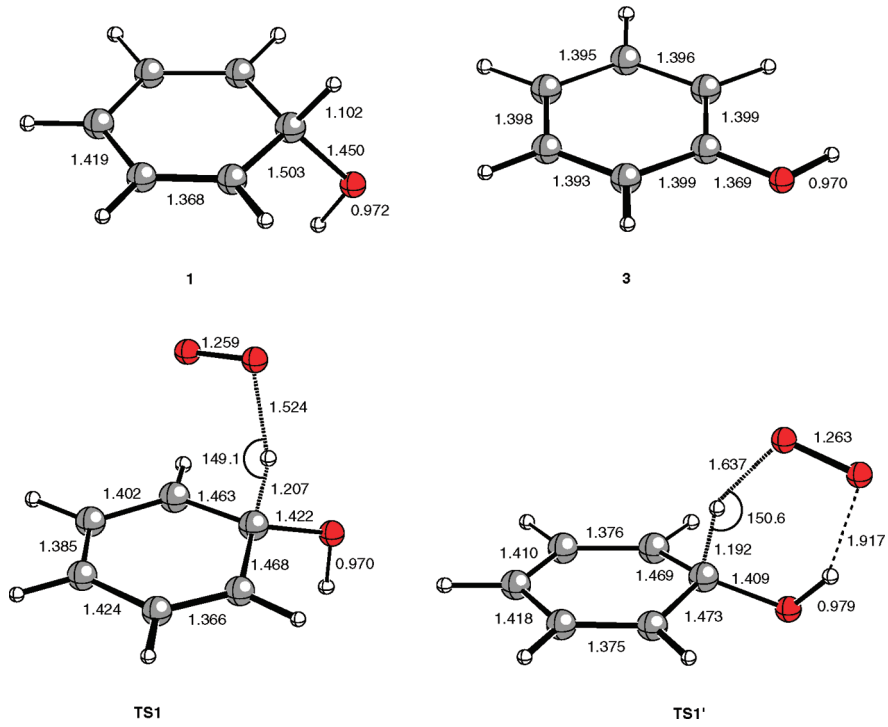


Figure 1. Selected geometrical parameters of the equilibrium structures of hydroxycyclohexadienyl radical (**1**), phenol (**3**), and the transition structures for the H-atom abstraction by O_2 in **1** affording **3**. Distances are given in angstroms and angles are in degrees.

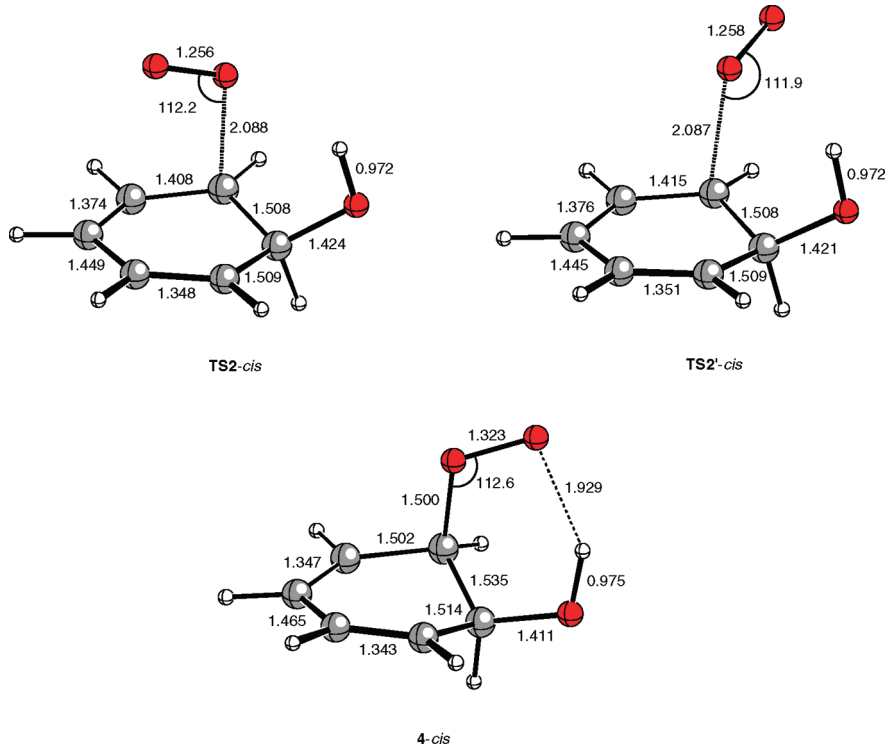


Figure 2. Selected geometrical parameters of the transition structures (**TS2**-*cis* and **TS2′**-*cis*) for the O_2 addition to position 2 of the benzene ring in radical **1** and the equilibrium structure (**4**-*cis*) of the cis stereoisomer of peroxyl radical **4**. Distances are given in angstroms and angles are in degrees.

hydrogen bond between these atoms. On the other hand, the AIM topological analysis of the electron density in **TS1** showed the presence of a bond critical point between one of the two oxygen atoms of the O_2 unit and the closer carbon atom at position 2 of the benzene ring, with an electron density of 0.0363 e. Therefore, although **TS1** does not show any hydrogen-bonding interaction, in this transition structure there exists an extra binding interaction between the O_2 molecule and the radical **1**, which is lacking in **TS1′**. Thus, the lower energy of **TS1** might be ascribed to the larger value of the electron density

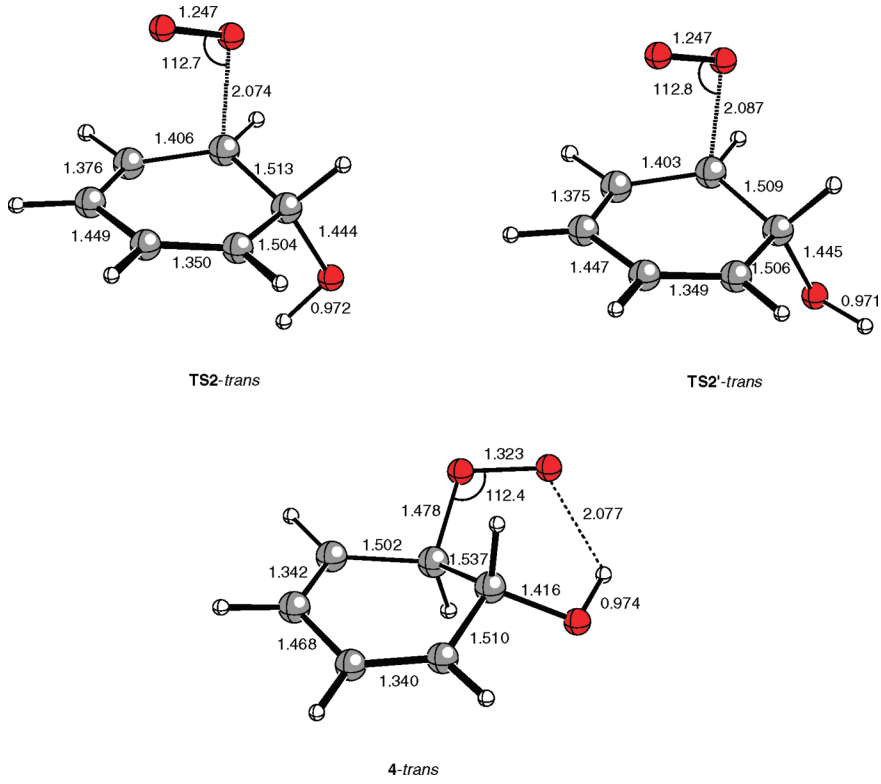


Figure 3. Selected geometrical parameters of the transition structures (**TS2**-*trans* and **TS2′**-*trans*) for the O_2 addition to position 2 of the benzene ring in radical **1** and the equilibrium structure (**4**-*trans*) of the trans stereoisomer of peroxyl radical **4**. Distances are given in angstroms and angles are in degrees.

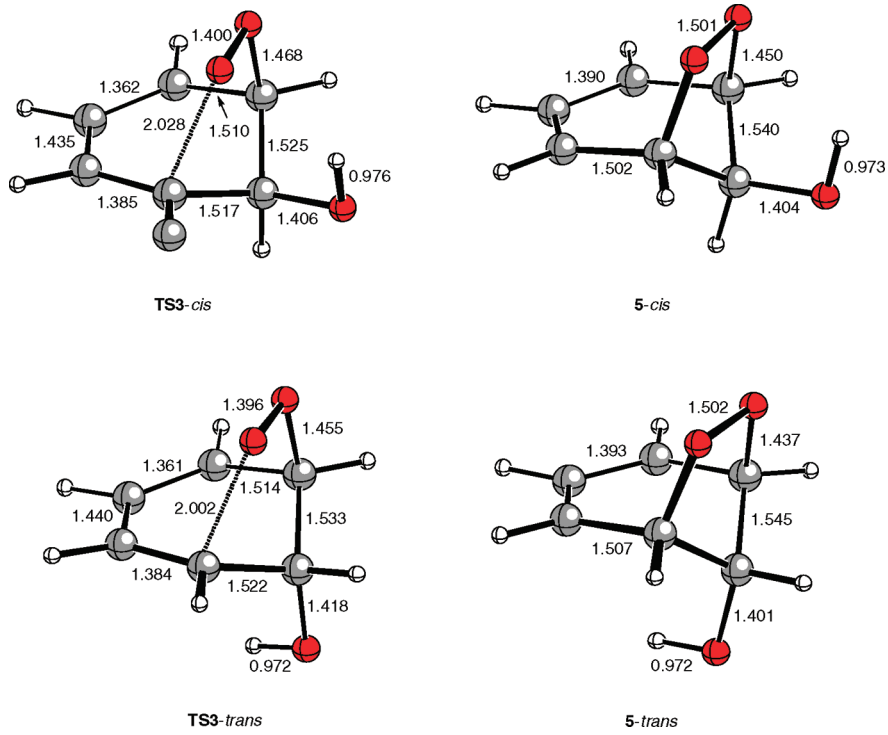


Figure 4. Selected geometrical parameters of the transition structures (**TS3**-*cis* and **TS3**-*trans*) for the cyclization of radicals **4**-*cis* and **4**-*trans* to bicyclic allyl radical **5** and the equilibrium structures (**5**-*cis* and **5**-*trans*) of the cis and trans stereoisomers of this radical. Distances are given in angstroms.

calculated at the bond critical point associated with the latter binding interaction in **TS1**, as compared to that calculated at the bond critical point associated with the hydrogen-bonding interaction in **TS1′**.

At the UCCSD(T) and RCCSD(T) levels, the energy difference between **TS1** and **TS1′** increases to the values of 3.0 and 5.9 kcal/mol, respectively. The $\langle S^2 \rangle$ values calculated for the UHF/6-311+G(2df,2p) wave function of **TS1** and

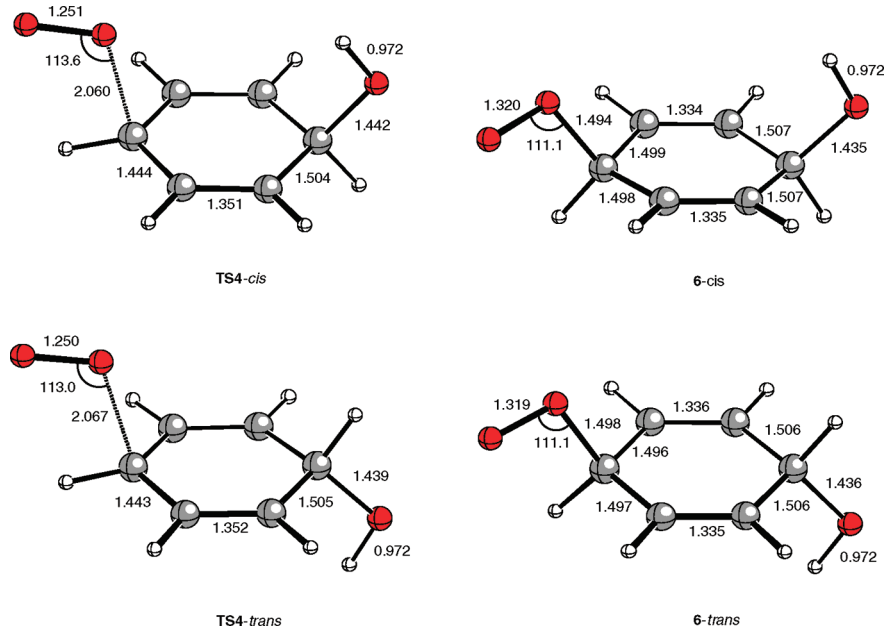


Figure 5. Selected geometrical parameters of the transition structures (**TS4**-*cis* and **TS4**-*trans*) for the O_2 addition to position 4 of the benzene ring in radical **1** and the equilibrium structures (**6**-*cis* and **6**-*trans*) of the cis and trans stereoisomers of peroxyl radical **6**. Distances are given in angstroms and angles are in degrees.

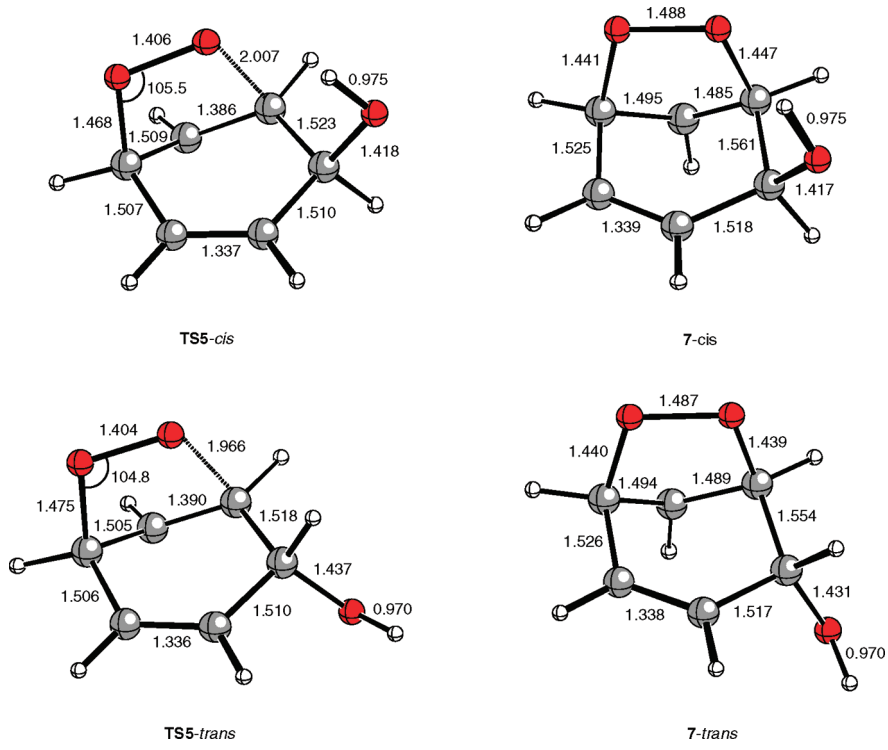


Figure 6. Selected geometrical parameters of the transition structures (**TS5**-*cis* and **TS5**-*trans*) for the cyclization of radicals **6**-*cis* and **6**-*trans* to bicyclic allyl radical **7** and the equilibrium structures (**7**-*cis* and **7**-*trans*) of the cis and trans stereoisomers of this radical. Distances are given in angstroms and angles are in degrees.

TS1′ are 2.08 and 2.19, respectively. As a consequence, the heights of the barriers computed with the UCCSD(T) and RCCSD(T) methods should differ significantly. In fact, focusing on the lowest energy transition structure **TS1**, the energy barrier in terms of ΔU (designated by $\Delta U^{\ddagger}$) calculated with these methods is 11.8 and 4.7 kcal/mol, respectively (see Table 1). The energy barrier calculated at the RCCSD(T) level leads to a rate coefficient at 298 K of 4.2×10^{-18} molecule^{-1} cm^3 s^{-1} (see Table 6), which is a factor of about 10 lower than the estimated experimental[22] value of $(6-11) \times 10^{-17}$ molecule^{-1} cm^3 s^{-1}. On the other hand, the energy barrier calculated at the UCCSD(T) level leads to a rate coefficient at 298 K of 3.1×10^{-23} molecule^{-1} cm^3 s^{-1}, which is too low by a factor of 10^6 as compared to the estimated experimental result. Therefore, it appears that the RCCSD(T) method performs much better than UCCSD(T) in the calculation of the energy barrier for H-atom abstraction from **1** by O_2 affording phenol. This finding is consistent

Table 1. Relative Energies (kcal/mol) of the Most Relevant Stationary Points on the Ground-State Potential Energy Surface for H-Atom Abstraction by O_2 in Hydroxycyclohexadienyl Radical (**1**) Forming Phenol (**3**)

	UB3LYP[a]	RCCSD(T)[b]			
stationary point[c]	ΔU	ΔU	ΔE(0 K)	ΔH(298 K)	ΔG(298 K)
1 + O_2	0.0	0.0	0.0	0.0	0.0
TS1	3.0	4.7 (11.8)	4.5 (11.5)	3.7 (10.8)	14.7 (21.8)
TS1′	4.4	10.6 (14.8)	9.8 (14.0)	9.2 (13.4)	19.3 (23.5)
3 + HOO$^{\bullet}$	−26.2	−29.9 (−30.3)	−29.2 (−29.6)	−29.3 (−30.9)	−29.6 (−30.0)

[a] Calculated with the 6-31G(d) basis set. [b] Calculated with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses. [c] See Figure 1.

Table 2. Relative Energies (kcal/mol) of the Most Relevant Stationary Points on the Ground-State Potential Energy Surface for O_2 Addition to Position 2 of Benzene Ring in Hydroxycyclohexadienyl Radical (**1**)

	UB3LYP[a]	RCCSD(T)[b]			
stationary point[c]	ΔU	ΔU	ΔE(0 K)	ΔH(298 K)	ΔG(298 K)
1 + O_2	0.0	0.0	0.0	0.0	0.0
TS2-*cis*	2.3	1.5 (6.6)	3.3 (8.4)	2.5 (7.6)	13.9 (19.0)
TS2′-*cis*	4.8	7.0 (10.8)	8.1 (12.0)	7.5 (11.4)	18.0 (21.8)
4-*cis*	−8.3	−12.7 (−13.4)	−9.4 (−10.1)	−10.3 (−11.0)	1.4 (0.7)
TS2-*trans*	3.1	1.5 (6.4)	3.2 (8.1)	2.5 (7.4)	13.6 (18.4)
TS2′-*trans*	5.9	3.9 (8.8)	5.2 (10.1)	4.7 (9.6)	15.3 (20.2)
4-*trans*	−9.4	−14.0 (−14.7)	−10.6 (−11.3)	−11.5 (−12.2)	0.0 (−0.7)

[a] Calculated with the 6-31G(d) basis set. [b] Calculated with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses. [c] See Figures 2 and 3.

Table 3. Relative Energies (kcal/mol) of the Most Relevant Stationary Points on the Ground-State Potential Energy Surface for the Cyclization of Peroxyl Radical **4** to the Bicyclic Allyl Radical **5**

	UB3LYP[a]	RCCSD(T)[b]			
stationary point[c]	ΔU	ΔU	ΔE(0 K)	ΔH(298 K)	ΔG(298 K)
4-*cis*	0.0	0.0	0.0	0.0	0.0
TS3-*cis*	10.6	13.4 (13.6)	13.1 (13.3)	12.4 (12.7)	14.1 (14.3)
5-*cis*	−8.1	−11.7 (−10.8)	−11.4 (−10.5)	−11.9 (−11.0)	−10.6 (−9.7)
4-*trans*	0.0	0.0	0.0	0.0	0.0
TS3-*trans*	17.1	18.8 (19.4)	18.4 (19.0)	17.9 (18.5)	19.5 (20.0)
5-*trans*	−3.8	−7.0 (−6.2)	−6.7 (−5.9)	−7.2 (−6.4)	−5.7 (−4.9)

[a] Calculated with the 6-31G(d) basis set. [b] Calculated with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses. [c] See Figures 2−4.

Table 4. Relative Energies (kcal/mol) of the Most Relevant Stationary Points on the Ground-State Potential Energy Surface for O_2 Addition to Position 4 of Benzene Ring in Hydroxycyclohexadienyl Radical (**1**)

	UB3LYP[a]	RCCSD(T)[b]			
stationary point[c]	ΔU	ΔU	ΔE(0 K)	ΔH(298 K)	ΔG(298 K)
1 + O_2	0.0	0.0	0.0	0.0	0.0
TS4-*cis*	4.2	4.6 (8.1)	6.0 (9.5)	5.5 (9.0)	15.2 (18.7)
6-*cis*	−8.3	−12.4 (−13.0)	−9.0 (−9.6)	−9.8 (−10.4)	1.0 (0.4)
TS4-*trans*	4.0	4.8 (8.2)	6.1 (9.6)	5.7 (9.1)	15.5 (18.9)
6-*trans*	−7.0	−12.3 (−12.9)	−9.0 (−9.6)	−11.7 (−12.3)	−0.9 (−1.5)

[a] Calculated with the 6-31G(d) basis set. [b] Calculated with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses. [c] See Figure 5.

with the fact that spin contamination is eliminated in the RCCSD(T) calculation of $\Delta U^{\ddagger}$.

We note that the values (1.18 and 1.19) of the tunneling factor Γ obtained with the RCCSD(T) and UCCSD(T) methods for the phenol channel (see Table 6) indicate that the tunneling effect in this reaction is negligible. This feature is contrary to common belief that for an H-atom transfer process the tunneling effect should be important. This unexpected result is ascribed to the fact that the energy barriers of reaction pathway **A** in Scheme 2 are broad, as suggested by the small value (415.9i cm^{-1}) of the imaginary vibrational frequency of the transition structure **TS1**.

To compare the results of our computations for the phenol channel with those of Lesclaux and co-workers,[22] it is convenient to consider the calculated enthalpy of activation at 298 K (designated by $\Delta H^{\ddagger}$(298 K)). By using a combination of UB3LYP/6-31G(d) and UCCSD(T)/6-31G(d,p) computations with an empirical relationship between $\Delta H^{\ddagger}$(298 K) and the reaction enthalpy at 298 K (designated by ΔH_r(298 K)), Lesclaux and co-workers[22] determined a $\Delta H^{\ddagger}$(298 K) of 1.8 kcal/mol. This value leads to a rate coefficient at 298 K of 1.4×10^{-16} molecule^{-1} cm^3 s^{-1}, which is in good agreement with the estimated experimental value. On the other hand, the RCCSD(T)/6-311+G(2df,2p)

Table 5. Relative Energies (kcal/mol) of the Most Relevant Stationary Points on the Ground-State Potential Energy Surface for the Cyclization of Peroxyl Radical **6** to the Bicyclic Allyl Radical **7**

	UB3LYP[a]	RCCSD(T)[b]			
stationary point[c]	ΔU	ΔU	ΔE(0 K)	ΔH(298 K)	ΔG(298 K)
6-*cis*	0.0	0.0	0.0	0.0	0.0
TS5-*cis*	30.3	30.7 (31.6)	29.9 (30.7)	29.1 (29.8)	31.6 (32.4)
7-*cis*	11.0	7.6 (8.1)	7.8 (8.3)	7.1 (7.6)	9.4 (9.9)
6-*trans*	0.0	0.0	0.0	0.0	0.0
TS5-*trans*	35.4	34.2 (34.9)	32.7 (33.5)	34.2 (35.0)	36.0 (36.7)
7-*trans*	15.5	10.8 (11.4)	10.6 (11.2)	12.0 (12.6)	13.8 (14.4)

[a] Calculated with the 6-31G(d) basis set. [b] Calculated with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses. [c] See Figures 5 and 6.

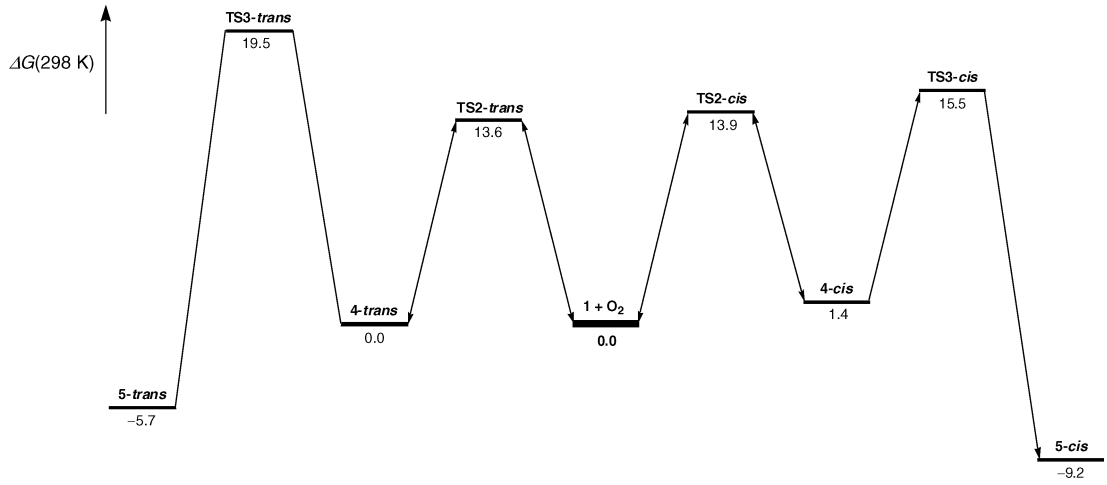


Figure 7. Schematic Gibbs energy profiles of the relevant reaction pathways concerning the O_2 addition to position 2 of the benzene ring in radical **1** and the subsequent ring closure of the peroxyl radicals **4**-*cis* and **4**-*trans*. Relative Gibbs energy values at 298 K (ΔG(298 K)) determined from RCCSD(T)/6-311+G(2df,2p) calculations.

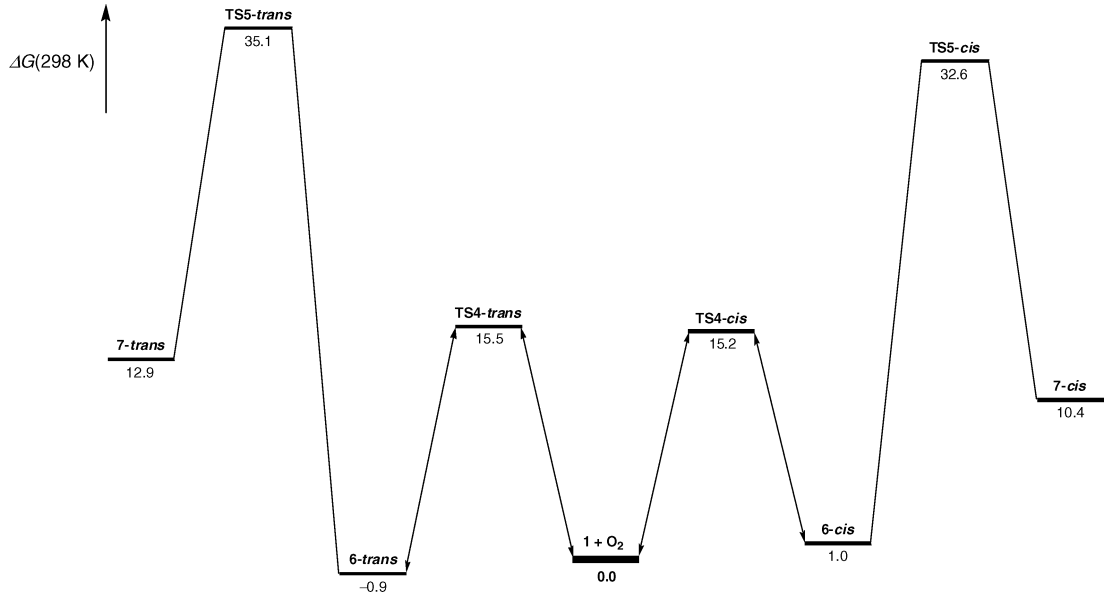


Figure 8. Schematic Gibbs energy profiles of the relevant reaction pathways concerning the O_2 addition to positions 4 of the benzene ring in radical **1** and the subsequent ring closure of the peroxyl radicals **6**-*cis* and **6**-*trans*. Relative Gibbs energy values at 298 K (ΔG(298 K)) determined from RCCSD(T)/6-311+G(2df,2p) calculations.

calculations predicts a $\Delta H^{\ddagger}$(298 K) of 3.7 kcal/mol (see Table 1), which is 1.9 kcal/mol higher than the value obtained by Lesclaux and co-workers and leads to a rate coefficient at 298 K that is a factor of about 10 lower than the estimated experimental value. Therefore, one might think that the semiempirical procedure of Lesclaux and co-workers performs much better than our approach, based on RCCSD(T)/6-311+G(2df,2p) calculations, in predicting activation energy barriers. However, our approach has the advantage of avoiding the use of empirical relationships between $\Delta H^{\ddagger}(T)$

Table 6. Tunneling Factor (Γ; See eq 2) and Rate Coefficient (k in molecule^{-1} cm^3 s^{-1}; See eq 1) at 298 K Calculated for Bimolecular Reactions Forming Phenol (**3**) and Peroxyl Radicals **4** and **6**[a]

reaction	TS	Γ	k
1 + O_2 → **3** + HOO•	**TS1**	1.18 (1.19)	4.2×10^{-18} (3.1×10^{-23})
1 + O_2 → **4**-*cis*	**TS2**-*cis*	1.19 (1.12)	1.9×10^{-17} (3.4×10^{-21})
1 + O_2 → **4**-*trans*	**TS2**-*trans*	1.19 (1.12)	3.2×10^{-17} (9.2×10^{-21})
1 + O_2 → **6**-*cis*	**TS4**-*cis*	1.17 (1.15)	2.1×10^{-18} (5.7×10^{-21})
1 + O_2 → **6**-*trans*	**TS4**-*trans*	1.16 (1.15)	1.3×10^{-18} (4.0×10^{-21})

[a] Calculated at the RCCSD(T) level of theory with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses.

Table 7. Tunneling Factor (Γ; See eq 2) and Rate Coefficient (k in s^{-1}; See eq 1) at 298 K Calculated for Unimolecular Reactions of Cyclization of Peroxyl Radicals **4** and **6** and Their Reversible Decomposition Back to **1** + O_2[a]

reaction	TS	Γ	k
4-*cis* → **5**-*cis*	**TS3**-*cis*	1.26 (1.12)	3.6×10^{2} (2.6×10^{2})
4-*cis* → **1** + O_2	**TS3**-*cis*	1.19 (1.12)	5.1×10^{3} (2.7×10^{-1})
4-*trans* → **5**-*trans*	**TS3**-*trans*	1.31 (1.31)	4.1×10^{-2} (1.8×10^{-2})
4-*trans* → **1** + O_2	**TS3**-*trans*	1.19 (1.12)	7.9×10^{2} (7.0×10^{-2})
6-*cis* → **7**-*cis*	**TS5**-*cis*	1.79 (1.80)	7.7×10^{-11} (2.0×10^{-11})
6-*cis* → **1** + O_2	**TS5**-*cis*	1.17 (1.15)	2.8×10^{2} (2.8×10^{-1})
6-*trans* → **7**-*trans*	**TS5**-*trans*	1.91 (1.92)	4.9×10^{-14} (1.5×10^{-14})
6-*trans* → **1** + O_2	**TS5**-*trans*	1.16 (1.159)	6.8×10^{1} (7.9×10^{-3})

[a] Calculated at the RCCSD(T) level of theory with the 6-311+G(2df,2p) basis set. The values calculated at the UCCSD(T) level with the same basis set are given in parentheses.

and $\Delta H_r(T)$. Furthermore, as shown below, the relative energy barriers determined at the RCCSD(T) level of theory with the 6-311+G(2df,2p) basis set for the competing reactions arising from the reaction **1** + O_2 turn out to be reasonably reliable.

Regarding the ΔH_r(298 K) of the phenol channel (see Table 1), it is to be noted that the value of −30.9 kcal/mol calculated with the UCCSD(T) method using the 6-311+G(2df,2p) basis set is 4.4 kcal/mol more negative than the value of −26.5 kcal/mol calculated with the same method using the 6-31G(d,p) basis set, reported by Lesclaux and co-workers.[22] It turns out, therefore, that the ΔH_r(298 K) calculated for the reaction channel affording phenol depends significantly on the size of the basis set employed. Furthermore, there is a difference of 1.6 kcal/mol between the values of ΔH_r(298 K) calculated with the UCCSD(T) and RCCSD(T) methods using the 6-311+G(2df,2p) basis set. This feature is ascribed to the significant spin contamination in the UHF/6-311+G(2df,2p) wave function of **1** (i.e., $\langle S^2 \rangle = 1.17$).

Earlier energy calculations by Ghigo and Tonachini[23,26] concerning the phenol channel, performed at the single-point UB3LYP/6-311+G(d,p) level using UB3LYP/6-31(+)G(d)-optimized geometries with energies corrected for spin contamination, are in reasonable agreement with the results of our RCCSD(T) calculations. Both the energy barrier and the energy of reaction in terms of ΔG(298 K) reported by Ghigo and Tonachini[23] (i.e., 13.6 and −27.6 kcal/mol, respectively) compare fairly well with our values of 14.7 and −29.6 kcal/mol, obtained at the RCCSD(T)/6-311+G(2df,2p) level.

3.2. O_2 Addition to Position 2 of the Benzene Ring in Hydroxycyclohexadienyl Radical. Table 2 gives the values of ΔU, ΔE(0 K), ΔH(298 K), and ΔG(298 K) calculated at different levels of theory for the relevant stationary points of the reaction pathway B in Scheme 2. In agreement with the results of earlier theoretical studies,[20,22] we found that the lowest energy structure of both the cis and the trans stereoisomers of peroxyl radical **4** (labeled as **4**-*cis* and **4**-*trans* in Figures 2 and 3, respectively) shows an intramolecular hydrogen bond between the terminal oxygen atom of the O−O fragment and the H-atom of the OH group. The lowest energy isomer corresponds to **4**-*trans*, whose ΔH(298 K) calculated at either the UCCSD(T) or the RCCSD(T) level of theory is 1.2 kcal/mol lower than that of **4**-*cis*. The ΔH_r(298 K) values determined with the UCCSD(T) method for the reactions affording **4**-*cis* and **4**-*trans* (i.e., −11.0 and −12.2 kcal/mol, respectively) are 0.7 kcal/mol more negative than those calculated with the RCCSD(T) (i.e., −10.3 and −11.5 kcal/mol, respectively). Lesclaux and co-workers[22] have reported an experimental ΔH_r(298 K) value of −12.5 kcal/mol for the O_2 addition to position 2 of the benzene ring in **1** affording peroxyl radical **4**. This value was determined from the measured thermodynamic equilibrium constant at 295 K for this reaction, assuming that the observed equilibrium must essentially involve the trans isomer of **4**, by using a reaction entropy at 298 K of −38.6 cal/mol·K (obtained from the UB3LYP/6-31G(d) calculations). At this point, we note that Lesclaux and co-workers assume that the chemical equilibrium **1** + O_2 ↔ **4** essentially involves the trans isomer of **4** on the basis of theoretical calculations predicting that the cis isomer is energetically less stable than the trans one and that its formation rate is significantly slower. However, it is not possible to resolve these isomers on the basis of the currently available experimental data. In the processing of the temperature-dependent equilibrium constants experimental data, it was assumed a single isomer of radical **4**, hence not differentiating the cis and trans isomers.[22]

The equilibrium between formation and decomposition of the peroxyl radicals **4**-*cis* and **4**-*trans* was evaluated according to eqs 3 and 4. The equilibrium constants at 298 K predicted by the RCCSD(T) and UCCSD(T) methods are given in Table 8. The equilibrium constant of reaction pathway **B** in Scheme 2 has been measured in several experimental studies.[18,20−22] Bohn and Zetzsch[18] determined an equilibrium constant of 2.7×10^{-19} cm^3 molecule^{-1} at 298 K, whereas Lesclaux and co-workers reported a value of 1.15×10^{-19} cm^3 molecule^{-1} at 295 K. More recently,

Table 8. Standard Gibbs Energy Change at 1 atm and 298 K (ΔG° in kcal mol^{-1}) and Equilibrium Constants (K_c in molecule^{-1} cm^3) at 298 K Calculated for the Reactions Forming Peroxyl Radicals **4** and **6**

	RCCSD(T)[a]		UCCSD(T)[b]	
reaction	ΔG°	K_c	ΔG°	K_c
1 + O_2 ↔ **4**-*cis*	1.4	3.82×10^{-21}	0.7	1.25×10^{-20}
1 + O_2 ↔ **4**-*trans*	0.0	4.06×10^{-20}	−0.7	1.32×10^{-19}
1 + O_2 ↔ **6**-*cis*	1.0	7.51×10^{-21}	0.4	2.07×10^{-20}
1 + O_2 ↔ **6**-*trans*	−0.9	1.86×10^{-19}	−1.5	5.11×10^{-19}

[a] Determined from relative energies calculated at the RCCSD(T) level of theory with the 6-311+G(2df,2p) basis set. [b] Determined from relative energies calculated at the UCCSD(T) level of theory with the 6-311+G(2df,2p) basis set.

the latter authors measured a value of $(2.62 \pm 0.24) \times 10^{-19}$ cm^3 molecule^{-1} at 295 K, which is in good agreement with that reported by Bohn and Zetzsch. Table 8 shows that the equilibrium constants for **4**-*cis* and **4**-*trans* obtained from the RCCSD(T) calculations are about a factor of 71 and 7, respectively, lower than the experimental value reported by Bohn and Zetzsch.[18] Interestingly, the values determined from the UCCSD(T) calculations for **4**-*cis* and **4**-*trans* are about a factor of 22 and 2, respectively, lower than the latter experimental value. These results suggest that the values of ΔH_r(298 K) predicted by the UCCSD(T) method are more reliable than those predicted by RCCSD(T) method. However, it should be stressed again that the experiments performed in all of these studies could not distinguish between the possible isomers of peroxyl radical **4**. This fact should be taken into account when comparing the experimental results with the theoretical calculations.

We found two transition structures for the reaction channel leading to **4**-*cis* (labeled as **TS2**-*cis* and **TS2′**-*cis* in Figure 2) and two transition structures for the reaction channel leading to **4**-*trans* (labeled as **TS2**-*trans* and **TS2′**-*trans* in Figure 3). In the case of **TS2**-*cis* and **TS2′**-*cis*, their geometries differ one from the other essentially in the orientation of the O−O bond relative to the benzene cycle, **TS2**-*cis* with the O−O bond nearly eclipsing a C−C bond and **TS2′**-*cis* with the O−O bond pointing away from the ring. The geometry of **TS2**-*cis* is similar to that of the transition structure found at the UB3LYP/6-31(+)G(d) level reported in ref 26 (designated TS(B)). On the other hand, because the geometry of the transition structure found at the UB3LYP/6-31G(d) level for the O_2 addition to radical **1** affording the cis isomer of peroxyl radical **4** is not reported in ref 22, it is not possible to ascertain whether or not such a transition structure is identical to **TS2**-*cis*. However, the value of 5.0 kcal/mol calculated at UB3LYP/6-31G(d) level for the $\Delta H^\ddagger$(298 K) of this reaction pathway in ref 22 appears significantly higher than the value of 3.2 kcal/mol obtained at the same level of theory from data given in Table S4 (Supporting Information). It is likely that the transition structure found at the UB3LYP/6-31G(d) level for the O_2 addition to radical **1** yielding the cis isomer of peroxyl radical **4** in ref 22 corresponds to **TS2′**-*cis*.

The geometries of **TS2**-*trans* and **TS2′**-*trans* differ one from the other essentially in the orientation of the O−H bond relative to the benzene cycle, **TS2**-*trans* with the O−H bond pointing to the ring and **TS2′**-*trans* with the O−H bond pointing away from the ring. Earlier theoretical studies by Ghigo and Tonachini[23,26] on the **1** + O_2 → **4** reaction considered only the O_2 addition on the same side of the benzene ring as the OH group, affording the cis isomer of peroxyl radical **4**. Hence, the latter authors do not report any transition structure for the formation of the trans isomer of **4**. Furthermore, the geometry of the transition structure found for this reaction pathway by Lesclaux and co-workers is not reported in ref 22. Therefore, the geometries of either **TS2**-*trans* or **TS2′**-*trans* cannot be compared to that of any previously calculated transition structure.

The UB3LYP/6-31G(d) calculations predict that the total energy of **TS2**-*cis* is 2.5 kcal/mol lower than that of **TS2′**-*cis*, whereas the total energy of **TS2**-*trans* is 2.8 kcal/mol lower than that of **TS2′**-*trans* (see Table 2). At the UCCSD(T) level, these energy differences are found to be 4.2 and 2.4 kcal/mol, respectively. The $\langle S^2 \rangle$ values calculated for the UHF/6-311+G(2df,2p) wave function of **TS2**-*cis*, **TS2′**-*cis*, **TS2**-*trans*, and **TS2′**-*trans* are 2.04, 2.10, 2.04, and 2.06, respectively. As a consequence, the heights of the barriers computed with the UCCSD(T) and RCCSD(T) methods for the reaction channels affording either **4**-*cis* or **4**-*trans* should differ significantly. In fact, focusing on the lowest energy transition structure, the values of $\Delta U^\ddagger$ calculated with these methods are 6.6 and 1.5 kcal/mol (**TS2**-*cis*) and 6.4 and 1.5 kcal/mol (**TS2**-*trans*), respectively (see Table 2). The $\Delta U^\ddagger$ of 1.5 kcal/mol, calculated at the RCCSD(T) level for the reaction channels affording either **4**-*cis* or **4**-*trans*, leads to the rate coefficients at 298 K of 1.9×10^{-17} and 3.2×10^{-17} molecule^{-1} cm^3 s^{-1}, respectively (see Table 6). Thus, the RCCSD(T) calculations predict a global rate coefficient at 298 K of 5.1×10^{-17} molecule^{-1} cm^3 s^{-1} for the reaction yielding peroxyl radical **4**, which is a factor of about 15−25 lower than the experimental[21,22] values ranging between 7.7×10^{-16} and 13.1×10^{-16} molecule^{-1} cm^3 s^{-1}. On the other hand, the energy barriers of 6.6 and 6.4 kcal/mol, calculated at the UCCSD(T) level for the reaction channels affording **4**-*cis* and **4**-*trans*, respectively, lead to the rate coefficients at 298 K of 3.4×10^{-21} and 9.2×10^{-21} molecule^{-1} cm^3 s^{-1} (see Table 6). The UCCSD(T) calculations, therefore, predict a global rate coefficient at 298 K of 1.3×10^{-20} molecule^{-1} cm^3 s^{-1} for the reaction leading to peroxyl radical **4**, which is too low by a factor of 10^4−10^5 as compared to the experimental values. Thus, it appears that the RCCSD(T) method performs much better than the UCCSD(T) one in the calculation of the energy barriers for O_2 addition to position 2 of the benzene ring in **1** yielding peroxyl radical **4**. Again, this result is consistent with the fact that spin contamination is eliminated in the RCCSD(T) calculations.

At this point, it is worth noting that the rate coefficients at 298 K derived from the RCCSD(T) calculations indicate that the formation rate of **4**-*trans* is slightly faster (a factor of 1.7) than that of **4**-*cis* (see Table 6). This result is at variance with the theoretical calculations of Lesclaux and co-workers[22] predicting that the formation rate of the trans

isomer is substantially faster (a factor of 50) than that of the cis one. This important discrepancy is traced back to the lower energy of **TS2**-*cis* as compared to that of the transition structure found by Lesclaux and co-workers for the reaction channel affording the cis isomer of peroxyl radical **4**.

Earlier energy calculations by Ghigo and Tonachini[23] concerning the O_2 addition to radical **1** yielding the cis isomer of peroxyl radical **4**, performed at the single-point UB3LYP/6-311+G(d,p) level using UB3LYP/6-31(+)G(d)-optimized geometries with energies corrected for spin contamination, are at variance with the results of our RCCSD(T) computations. Ghigo and Tonachini reported, in terms of ΔG(298 K), an energy barrier of 15.6 kcal/mol and an energy of reaction of 10.5 kcal/mol, which are significantly higher than the values of 13.9 and 1.4 kcal/mol we obtained at the RCCSD(T)/6-311+G(2df,2p) level. The origin of such a large discrepancy in the calculated energy of reaction is unclear. It has been observed in similar systems that the B3LYP functional may yield significant differences in calculated energies of reaction, as compared to the values obtained from CCSD(T) calculations.[22,62] In general, CCSD(T) calculations describe the reactions as being more exoergic than do B3LYP calculations.

3.3. Cyclization Reaction of Peroxyl Radical 4. On the basis of previous theoretical calculations of thermochemical and kinetic parameters for the cyclization of peroxyl radical **4** affording bicyclic radicals by formation of a peroxy bridge, the ring closure in **4** leading to radical **5** (see Scheme 3) appears to be the only possible cyclization pathway for peroxyl radical **4** under tropospheric conditions.[23] Therefore, here we have considered only this cyclization mode for both **4**-*cis* and **4**-*trans*. Table 3 gives the values of ΔU, ΔE(0 K), ΔH(298 K), and ΔG(298 K) calculated at different levels of theory for the relevant stationary points associated with these cyclization reactions. In agreement with the results of the theoretical study by Lesclaux and co-workers,[22] the UB3LYP, UCCSD(T), and RCCSD(T) calculations predict the cis stereoisomer of the byciclic radical **5** (labeled as **5**-*cis* in Figure 4) to be energetically more stable than the trans stereoisomer (labeled as **5**-*trans* in Figure 4). For instance, the values of ΔH(298 K) determined from the UCCSD(T) and RCCSD(T) calculations for **5**-*cis* are 4.6 and 4.7 kcal/mol, respectively, lower than those calculated for **5**-*trans*.

The transition structures calculated for the cyclization of **4**-*cis* leading to **5**-*cis* (labeled as **TS3**-*cis*) and the cyclization of **4**-*trans* affording **5**-*trans* (labeled as **TS3**-*trans*) are depicted in Figure 4. Interestingly, the $\Delta U^{\ddagger}$ values computed with the UCCSD(T) and RCCSD(T) methods for these reactions differ only in a few tenths of kcal/mol. This finding might be ascribed to a small degree of spin contamination of the UHF wave function of **TS3**-*cis* and **TS3**-*trans*. However, the $\langle S^2 \rangle$ value calculated for the UHF/6-311+G(2df,2p) wave function of **TS3**-*cis* and **TS3**-*trans* is 1.33.

Earlier UB3LYP/6-311+G(d,p) calculations using UB3LYP/6-31(+)G(d)-optimized geometries with energies corrected for spin contamination by Ghigo and Tonachini[23] on the ring closure of peroxyl radical **4** affording the bicyclic radical **5** considered only the cis isomer of **4**. The $\Delta H^{\ddagger}$(298 K) of 12.7 kcal/mol reported by Ghigo and Tonachini for this reaction pathway is in good agreement with the value of 12.4 kcal/mol we obtained from our RCCSD(T)/6-311+G(2df,2p) calculations. On the contrary, the ΔH_r(298 K) of −5.4 kcal/mol computed by Ghigo and Tonachini differs substantially from the value of −11.9 kcal/mol of our RCCSD(T)/6-311+G(2df,2p) calculations. We recall again that on similar systems it has been observed that, in general, the CCSD(T) calculations predict the reactions as being more exoergic than do the B3LYP calculations.[22,62]

In line with the results reported by Lesclaux and co-workers,[22] all barrier heights ($\Delta U^{\ddagger}$, $\Delta E^{\ddagger}$(0 K), $\Delta H^{\ddagger}$(298 K), and $\Delta G^{\ddagger}$(298 K)) calculated for the cyclization reaction of **4**-*cis* are significantly lower than those calculated for the cyclization of **4**-*trans* (see Table 3). As a consequence, the value of the rate coefficient at 298 K derived from the RCCSD(T) calculations (see Table 7) for the cyclization **4**-*cis* → **5**-*cis* (3.6×10^2 s^{-1}) is a factor of about 10^4 higher than the value determined for the cyclization **4**-*trans* → **5**-*trans* (4.1×10^{-2} s^{-1}). On the other hand, Table 7 shows that the rate coefficient calculated for the reversible decomposition **4**-*cis* → **1** + O_2 (5.1×10^3 s^{-1}) is about a factor of 10 higher than rate coefficient obtained for the cyclization **4**-*cis* → **5**-*cis*, whereas the rate coefficient for the reversible decomposition **4**-*trans* → **1** + O_2 (7.9×10^2 s^{-1}) is about a factor of 10^4 higher than the rate coefficient for the cyclization **4**-*trans* → **5**-*trans*. Therefore, under tropospheric conditions, it appears that the only possible reaction pathway for **4**-*trans* is the reversible decomposition back to the reactants, leading to the chemical equilibrium **1** + O_2 ↔ **4**-*trans*, whereas **4**-*cis* can undergo cyclization to the bicyclic radical **5**-*cis*. These results are pictorially illustrated in Figure 7 in terms of the ΔG(298 K) calculated at the RCCSD(T) level of theory for the relevant stationary points involved in the O_2 addition to position 2 of the benzene ring in radical **1** and the subsequent ring closure of the peroxyl radicals formed. Because the bicyclic radical **5** can lead readily to cleavage of the former aromatic ring yielding the principal benzene oxidation products (glyoxal and butenedial),[24,25] it turns out that the formation of **4**-*cis* implies irreversible loss of radical **1**. On the other hand, the experimentally observed chemical equilibrium **1** + O_2 ↔ **4** must essentially involve the trans isomer of peroxyl radical **4**. This feature confirms the assumption put forward by Lesclaux and co-workers[22] on the basis that the trans isomer is energetically more stable and is formed much more rapidly (a factor of about 50) than the cis one. However, as emphasized above, the rate coefficients at 298 K derived from the RCCSD(T) calculations indicate that the formation rate of **4**-*trans* is only slightly faster (a factor of 1.7) than that of **4**-*cis* (see Table 6). Therefore, the observed chemical equilibrium **1** + O_2 ↔ **4** must essentially involve the trans isomer of **4** because the cyclization **4**-*trans* → **5**-*trans* cannot compete with the decomposition of **4**-*trans* back to the reactants **1** + O_2.

3.4. O_2 Addition to Position 4 of the Benzene Ring in Hydroxycyclohexadienyl Radical. Table 4 gives the values of ΔU, ΔE(0 K), ΔH(298 K), and ΔG(298 K) calculated at different levels of theory for the relevant stationary points of the reaction pathway shown in Scheme 4. Both the UCCSD(T) and the RCCSD(T) calculations predict the total energy of the cis stereoisomer of peroxyl radical **6** (labeled as **6**-*cis* in Figure 5) to be 0.1 kcal/mol lower than that of the trans one (labeled as **6**-*trans* in Figure 5). Inclusion of ZPVE and thermal corrections to energy changes the relative energy ordering of these isomers. Thus, the ΔH(298 K) calculated at either the UCCSD(T) or the RCCSD(T) level of theory for **6**-*trans* is 1.9 kcal/mol lower than that of **6**-*cis*. The ΔH_r(298 K) values determined with the UCCSD(T) method for the addition reactions affording **6**-*cis* and **6**-*trans* (i.e., −10.4 and −12.3 kcal/mol, respectively) are 0.6 kcal/mol more negative than those calculated with the RCCSD(T) (i.e., −9.8 and −11.7 kcal/mol, respectively). Interestingly, these ΔH_r(298 K) values differ only in a few tenths of kcal/mol from those calculated for the addition reactions affording **4**-*trans* and **4**-*cis* (compare the ΔH(298 K) values given in Tables 2 and 4). As a consequence, the values of the equilibrium constants predicted for the chemical equilibriums **1** + O_2 ↔ **6**-*cis* and **1** + O_2 ↔ **6**-*trans* are close to those predicted for the equilibriums **1** + O_2 ↔ **4**-*cis* and **1** + O_2 ↔ **4**-*trans*, respectively (see Table 8).

At variance with the addition reactions **1** + O_2 affording **4**-*cis* and **4**-*trans*, we found only one transition structure for the reaction channel leading to **6**-*cis* (labeled as **TS4**-*cis* in Figure 5) and only one transition structure for the reaction channel leading to **6**-*trans* (labeled as **TS4**-*trans* in Figure 5). The UB3LYP/6-31G(d) calculations predict that the total energy of **TS4**-*cis* is 0.2 kcal/mol higher than that of **TS4**-*trans* (see Table 4). At the UCCSD(T) level of theory, the total energy of **TS4**-*cis* is calculated to be 0.1 kcal/mol lower than that of **TS4**-*trans*. The $\langle S^2 \rangle$ values determined for the UHF/6-311+G(2df,2p) wave functions of **TS4**-*cis* and **TS4**-*trans* are 1.98 and 1.99, respectively. As a consequence, the heights of the barriers computed with the UCCSD(T) and RCCSD(T) methods for the reaction channels leading to **6**-*cis* and **6**-*trans* differ significantly. For instances, the values of $\Delta U^{\ddagger}$ calculated with the UCCSD(T) and RCCSD(T) methods for these reaction channels are 8.1 and 4.6 kcal/mol (**TS4**-*cis*) and 8.2 and 4.8 kcal/mol (**TS4**-*trans*), respectively (see Table 4).

Focusing on the energy barriers calculated with the RCCSD(T) method, a comparison between the O_2 addition to positions 2 and 4 of the benzene ring in radical **1** affording the cis and trans isomers of radicals **4** and **6** reveals that the $\Delta U^{\ddagger}$ for the addition to position 4 is about 3 kcal/mol higher than that for addition to position 2 (see Tables 2 and 4). The energy barriers determined from the RCCSD(T) calculations for the reaction channels affording **6**-*cis* and **6**-*trans* lead to the rate coefficients at 298 K of 2.1×10^{-18} and 1.3×10^{-18} molecule^{-1} cm^3 s^{-1}, respectively, which are a factor of about 10 lower than those calculated for the reaction channels affording **4**-*cis* and **4**-*trans* (see Table 6). Consequently, although the ΔH_r(298 K) values for the O_2 addition to positions 2 and 4 of the benzene ring in radical **1** are predicted to be similar, the addition to position 2 is clearly preferred over the addition to position 4.

3.5. Cyclization Reaction of Peroxyl Radical 6. From the strain energy point of view, the ring closure to the bicyclic radical **7** (see Scheme 5) appears to be the more viable cyclization mode of peroxyl radical **6** under tropospheric conditions. Therefore, here we have considered only this cyclization mode for both **6**-*cis* and **6**-*trans*. Table 5 gives the values of ΔU, ΔE(0 K), ΔH(298 K), and ΔG(298 K) calculated at different levels of theory for the relevant stationary points associated with these cyclization reactions. Both the UCCSD(T) and the RCCSD(T) calculations predict the cis stereoisomer of the bycyclic radical **7** (labeled as **7**-*cis* in Figure 6) to be energetically more stable than the trans stereoisomer (labeled as **7**-*trans* in Figure 6). Thus, the ΔH(298 K) values determined from the UCCSD(T) and RCCSD(T) calculations for **7**-*cis* are 5.0 and 4.9 kcal/mol lower, respectively, than those calculated for **7**-*trans*.

In clear contrast with the cyclization reactions **4**-*cis* → **5**-*cis* and **4**-*trans* → **5**-*trans*, which were found to be exothermic, the cyclizations **6**-*cis* → **7**-*cis* and **6**-*trans* → **7**-*trans* are calculated to be endothermic. For instance, the ΔH_r(298 K) values determined from the RCCSD(T) calculations for the reactions **4**-*cis* → **5**-*cis* and **4**-*trans* → **5**-*trans* are −11.9 and −7.2 kcal/mol, respectively (see Table 3), whereas those determined for the reactions **6**-*cis* → **7**-*cis* and **6**-*trans* → **7**-*trans* are 7.1 and 12.0 kcal/mol, respectively (see Table 5).

The transition structures found for the cyclization reactions **6**-*cis* → **7**-*cis* (labeled as **TS5**-*cis*) and **6**-*trans* → **7**-*trans* (labeled as **TS5**-*trans*) are shown in Figure 6. Although the $\langle S^2 \rangle$ values calculated for the UHF/6-311+G(2df,2p) wave function for **TS5**-*cis* (1.14) and **TS5**-*trans* (1.15) indicate a significant degree of spin contamination, the $\Delta U^{\ddagger}$ values computed with the UCCSD(T) and RCCSD(T) methods for these reactions are similar (see Table 5). As found for the cyclization reactions of **4**-*cis* and **4**-*trans*, all barrier heights ($\Delta U^{\ddagger}$, $\Delta E^{\ddagger}$(0 K), $\Delta H^{\ddagger}$(298 K), and $\Delta G^{\ddagger}$(298 K)) calculated for the cyclization of **6**-*cis* are significantly lower than those calculated for the cyclization of **6**-*trans* (see Table 5). However, all barrier heights computed for the cyclization of either **6**-*cis* or **6**-*trans* are about twice those computed for the cyclization of either **4**-*cis* or **4**-*trans* (see Tables 3 and 5). Furthermore, the values of the rate coefficient at 298 K derived from the RCCSD(T) calculations (see Table 7) for the cyclizations **6**-*cis* → **7**-*cis* (7.7×10^{-11} s^{-1}) and **6**-*trans* → **7**-*trans* (4.9×10^{-14} s^{-1}) are extremely small, as compared to those for the cyclizations **4**-*cis* → **5**-*cis* (3.6×10^{2} s^{-1}) and **4**-*trans* → **5**-*trans* (4.1×10^{-2} s^{-1}). On the other hand, Table 7 shows that the rate coefficient calculated for the reversible decomposition **6**-*cis* → **1** + O_2 (2.8×10^{2} s^{-1}) is about a factor of 10^{13} higher than the rate coefficient obtained for the cyclization **6**-*cis* → **7**-*cis* and the rate coefficient for the reversible decomposition **6**-*trans* → **1** + O_2 (6.8×10^{1} s^{-1}) is about a factor of 10^{15} higher than the rate coefficient for the cyclization **6**-*trans* → **7**-*trans*. Therefore, under tropospheric condi-

tions, it appears that the only possible reaction pathway for either **6**-*cis* or **6**-*trans* is the reversible decomposition back to the reactants, leading to the chemical equilibrium $\mathbf{1} + O_2 \leftrightarrow \mathbf{6}$. This feature is pictorially illustrated in Figure 8 in terms of the ΔG(298 K) calculated at the RCCSD(T) level of theory for the stationary points involved in the O_2 addition to position 4 of the benzene ring in radical **1** and the subsequent ring closure of the peroxyl radicals formed. As a consequence, it appears that the O_2 addition to position 4 of the benzene ring in radical **1** cannot contribute to the formation of benzene oxidation products through cleavage of the former aromatic ring in bicyclic radicals **7**-*cis* and **7**-*trans*.

3.6. Global Irreversible Loss of Hydroxycyclohexadienyl and Peroxyl Radicals. One of the main objectives of the experimental work carried out by Lesclaux and co-workers[22] on the reaction of radical **1** with O_2 was to provide kinetic data accounting for the global irreversible loss of radical species (essentially radical **1** and the resulting peroxyl radicals), yielding phenol and other oxidation products. Specifically, Lesclaux and co-workers measured experimentally a total rate coefficient for the global radical loss reactions of $(2.52 \pm 0.40) \times 10^{-16}$ cm^3 molecule^{-1} s^{-1} at 295 K. This rate coefficient was employed for evaluating product yields from calculated rate coefficients of possible reaction channels. In particular, from the rate coefficient at 298 K of 1.4×10^{-16} molecule^{-1} cm^3 s^{-1} calculated for the phenol channel, a yield of about 55% was obtained for phenol, in reasonable agreement with experimental values (25−61%).[1,30−32]

Besides the reaction channel yielding phenol, the theoretical results described above for the possible reaction channels arising from the reaction of radical **1** with O_2 indicate that the reaction channel $\mathbf{1} + O_2 \rightarrow$ **4**-*cis* is the only reaction leading to irreversible loss of radical **1**. Therefore, excluding radical−radical reactions, the total rate coefficient for the global irreversible loss of radicals species can be approximated as the sum of the rate coefficients at 298 K calculated for the reactions $\mathbf{1} + O_2 \rightarrow$ **3** + HOO$^\bullet$ and $\mathbf{1} + O_2 \rightarrow$ **4**-*cis*. This approximation leads to a total rate coefficient at 298 K of 2.3×10^{-17} molecule^{-1} cm^3 s^{-1} (see Table 6), which is a factor of about 10 lower than the experimental value of Lesclaux and co-workers. However, from the rate coefficient at 298 K of 4.2×10^{-18} molecule^{-1} cm^3 s^{-1} calculated for the phenol channel, a yield of about 18% was obtained for phenol, in reasonable agreement with experimental values.[1,30−32] Therefore, although the energy barriers obtained from RCCSD(T) calculations with the 6-311+G (2df,2p) basis set for the competing reaction channels arising from the reaction $\mathbf{1} + O_2$ lead to rate coefficients at 298 K that are a factor of about 10 too low, the relative rate coefficients are reasonably reliable.

3.7. Comparison to Theoretical Calculations on HO$^\bullet$-Initiated Oxidation of *p*-Xylene and *m*-Xylene. Recently, Fan and Zhang have studied the HO$^\bullet$-initiated oxidation reactions of *p*-xylene[63] and *m*-xylene.[64] By using optimized geometries, vibrational frequencies, and ZPVE-corrected energies, obtained at the UB3LYP/6-31G(d,p) level, Fan and Zhang have investigated the competing pathways arising from the reaction of the *p*-xylene−HO$^\bullet$ and *m*-xylene−HO$^\bullet$ adducts with O_2 to assess the energetically favorable pathways to propagate the oxidations. As compared to benzene oxidation, the mechanistic complexity of the *p*-xylene and *m*-xylene oxidations is much higher due to the existence of multiple isomeric pathways at each reaction stage.

The theoretical calculations of Fan and Zhang predict the HO$^\bullet$ addition to occur preferentially at the ortho position of *p*-xylene and the two possible ortho positions of *m*-xylene. Regarding the O_2 addition to the *p*-xylene−HO$^\bullet$ and *m*-xylene−HO$^\bullet$ adducts, the theoretical study of Fang and Zhang focuses exclusively on the addition on the same side of the benzene ring as the hydroxyl group, because they found that this addition mode leads to the formation of the energetically favorable isomers of the peroxyl radicals. In clear contrast, our UCCSD(T) and RCCSD(T) calculations predict that the peroxyl radical **4**-*trans* resulting from the O_2 addition to the benzene−HO$^\bullet$ adduct **1** is less energetic and is formed somewhat faster than the isomer **4**-*cis*.

The ZPVE-corrected reaction energies (designated by ΔE_r(0 K)) for the formation of HO$^\bullet$−*p*-xylene−O_2 and HO$^\bullet$−*m*-xylene−O_2 peroxyl radicals from the O_2 addition to the corresponding *p*-xylene−HO$^\bullet$ and *m*-xylene−HO$^\bullet$ adducts range from −4.5 to −7.1 kcal/mol. These values are significantly less negative than the ΔE_r(0 K) obtained from the UCCSD(T) and RCCSD(T) calculations (see Tables 2 and 4) for the reaction pathways $\mathbf{1} + O_2 \rightarrow$ **4**-*cis* (−10.1 and −9.4 kcal/mol, respectively) and $\mathbf{1} + O_2 \rightarrow$ **6**-*cis* (−9.6 and −9.0 kcal/mol, respectively). The ZPVE-corrected energy barriers (designated by $\Delta E^{\ddagger}$(0 K)) for the O_2 addition to the ortho *p*-xylene−HO$^\bullet$ adduct range from −0.51 to 4.18 kcal/mol, while for the O_2 addition to the two ortho *p*-xylene−HO$^\bullet$ adducts range from −1.2 to 3.56 kcal/mol. These barriers are lower than our $\Delta E^{\ddagger}$(0 K) values of 3.3 and 6.0 kcal/mol obtained with the RCCSD(T) method for the reaction pathways $\mathbf{1} + O_2 \rightarrow$ **4**-*cis* and $\mathbf{1} + O_2 \rightarrow$ **6**-*cis*, respectively.

Finally, it is worth noting that the ΔE_r(0 K) values reported by Fan and Zhang for the cyclization of the *p*-xylene and *m*-xylene peroxyl radicals arising from initial HO$^\bullet$ and subsequent O_2 addition to the ring to form bridged bicyclic radicals possessing a delocalized allyl system range between −5.37 and −8.89 kcal/mol. These values are less negative than the ΔE_r(0 K) obtained from the UCCSD(T) and RCCSD(T) calculations (see Table 3) for the isomerization of peroxyl radical **4**-*cis* to the bicyclic radical **5**-*cis* (−10.5 and −11.4 kcal/mol, respectively). Furthermore, the $\Delta E^{\ddagger}$(0 K) values reported by Fan and Zhang for the cyclization of HO$^\bullet$−*p*-xylene−O_2 and HO$^\bullet$−*m*-xylene−O_2 peroxyl radicals affording bridged bicyclic radicals containing a delocalized allyl radical (ranging from 9.07 to 11.14 kcal/mol) are significantly lower than the $\Delta E^{\ddagger}$(0 K) value of 13.1 kcal/mol predicted by the RCCSD(T) calculations for the isomerization of peroxyl radical **4**-*cis* to the bicyclic radical **5**-*cis*.

4. Summary and Conclusions

Density functional theory (UB3LYP) and quantum-mechanical (UCCSD(T) and RCCSD(T)) electronic structure calculations were carried out to investigate the primary steps of the oxidative degradation of benzene under tropospheric conditions, initiated by the addition of HO• to the aromatic ring. The energetic, structural, and vibrational results furnished by these calculations were subsequently used to perform conventional transition-state computations to predict the rate coefficients and evaluate the product yields of the competing abstraction and addition reactions arising from the reaction of the benzene−HO• adduct **1** with O_2. From the analysis of the results, the following main points emerge.

(1) The barrier heights ($\Delta U^{\ddagger}$, $\Delta E^{\ddagger}$, $\Delta H^{\ddagger}$, and $\Delta G^{\ddagger}$) determined from RCCSD(T) calculations with the 6-311+G(2df,2p) basis set are found to be more reliable than those obtained from UCCSD(T) calculations with the same basis set. This theoretical finding is ascribed to the high degree of spin contamination shown by the UHF wave function underlying the UCCSD(T) calculations of the transition structures and is consistent with the fact that such spin contamination is eliminated in the RCCSD(T) calculations.

(2) It is confirmed that the trans stereoisomer of the peroxyl radical **4** produced by the O_2 addition to position 2 of benzene ring in the benzene−HO• adduct **1** is energetically more stable than the cis one. However, at variance with an earlier theoretical study, the rate coefficients at 298 K for the formation of both stereoisomers are predicted to be similar.

(3) All of the barrier heights ($\Delta U^{\ddagger}$, $\Delta E^{\ddagger}$, $\Delta H^{\ddagger}$, and $\Delta G^{\ddagger}$) calculated for the cyclization of the cis isomer of peroxyl radical **4** to the cis isomer of a bicyclic allyl radical **5** bearing a peroxy bridge are significantly lower than those calculated for the cyclization of the trans isomer of **4**. Because radical **5** can lead readily to cleavage of the former aromatic ring yielding the principal benzene oxidation products, it is concluded that the formation of the cis isomer of **4** implies irreversible loss of radical **1** and that the observed chemical equilibrium **1** + O_2 ↔ **4** must essentially involve the trans isomer of **4**.

(4) The O_2 addition to position 4 of benzene ring in the benzene−HO• adduct **1** affords the cis and trans stereoisomers of a peroxyl radical **6**. The cis isomer of **6** is predicted to be energetically more stable than the trans one. Although the reaction enthalpies calculated for the O_2 addition to positions 2 and 4 of the benzene ring in radical **1** are calculated to be similar, the addition to position 2 is clearly preferred over the addition to position 4 because it involves a lower barrier.

(5) The heights of the barriers computed for the cyclization of either the cis or the trans isomer of peroxyl radical **6** to a bicyclic radical **7** bearing a peroxy bridge are about twice the heights of the barriers computed for the cyclization of either the cis or the trans isomer of peroxyl radical **4** to bicyclic radical **5**. Under tropospheric conditions, the only possible reaction pathway for radical **6** is the reversible decomposition back to the reactants, leading to the chemical equilibrium **1** + O_2 ↔ **6**. As a consequence, it is unlikely that the O_2 addition to position 4 of the benzene ring in radical **1** can contribute to the formation of benzene oxidation products.

Acknowledgment. This research was supported by the Spanish MEC (Grants CTQ2005-07790-C02-01 and CTQ2005-01117). Additional support came from the Catalonian AGAUR (Grants 2005SGR00111 and 2005PEIR0051/69). The larger calculations described in this work were performed at the Centre de Supercomputació de Catalunya (CESCA). The reviewers provided helpful comments for improving this Article.

Supporting Information Available: The $\langle S^2 \rangle$ values of the UHF/6-311+G(2df,2p) wave functions, open-shell T_1 diagnostic values of the RCCSD calculations, total energies, zero-point vibrational energies, and thermal corrections to enthalpy and Gibbs energy, as well as the Cartesian coordinates of all structures reported in this Article. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Calvert, J. G.; Atkinson, R.; Becker, K. H.; Kamens, R. M.; Seinfeld, J. H.; Wallington, T. J.; Yarwood, G. *The Mechanism of Atmospheric Oxidation of Aromatic Hydrocarbons*; Oxford University Press: New York, 2002.

(2) Piccot, S. D.; Watson, J. J.; Jones, J. W. *J. Geophys. Res.* **1992**, *97*, 9897.

(3) Legett, S. *Atmos. Environ.* **1996**, *30*, 215.

(4) Derwent, R. G.; Jenkin, M. E.; Saunders, S. M. *Atmos. Environ.* **1996**, *30*, 181.

(5) Seinfeld, J. H.; Pandis, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; John Wiley & Sons: New York, 1997.

(6) Lay, T. H.; Bozzelli, J. W.; Seinfeld, J. H. *J. Phys. Chem.* **1996**, *100*, 6543.

(7) Atkinson, R.; Aschmann, S. M. *Int. J. Chem. Kinet.* **1994**, *26*, 929.

(8) Atkinson, R.; Aschmann, S. M.; Arey, J.; Carter, W. P. L. *Int. J. Chem. Kinet.* **1989**, *21*, 801.

(9) Semadeni, M.; Stocker, D. W.; Kerr, J. A. *Int. J. Chem. Kinet.* **1995**, *27*, 287.

(10) Wallington, T. J.; Neuman, D. J.; Kurylo, M. J. *Int. J. Chem. Kinet.* **1987**, *19*, 725.

(11) Witte, F.; Urbanik, E.; Zetzsch, C. *J. Phys. Chem.* **1986**, *90*, 3251.

(12) Knispel, R.; Koch, R.; Seise, M.; Zetzsch, C. *Ber. Bunsen-Ges. Phys. Chem.* **1990**, *94*, 1375.

(13) Lin, S.-C.; Kuo, T.-C.; Lee, Y.-P. *J. Chem. Phys.* **1994**, *101*, 2098.

(14) Bjergbakke, E.; Sillesen, A.; Pagsberg, P. J. *J. Phys. Chem.* **1996**, *100*, 5729.

(15) Zellner, R.; Fritz, B.; Priedel, M. *Chem. Phys. Lett.* **1985**, *121*, 412.

(16) Zetzsh, C.; Koch, R.; Bohn, B.; Knispel, R.; Siese, M.; Witte, F. In *Transp. Chem. Transform. Pollut. Troposphere*; Le

Bras, G., Ed.; Springer: Berlin, Germany, 1997; Vol. 3, pp 347−356.

(17) Approximate atmospheric concentrations of the reactive species can be found, for example, in: (a) Güsten, H. Degradation of Atmospheric Pollutants by Tropospheric Free Radical Reactions. In *Free Radicals in Biology and Environment*; Minisci, F., Ed.; Kluwer Academic Publishers: The Netherlands, 1997; Chapter 28, pp 387−408. (b) Atkinson, R. Reactions of Oxygen Species in the Atmosphere. In *Active Oxygen in Chemistry*; Valentine, J. S., Foote, C. S., Greenberg, A., Liebman, J. F., Eds.; Blackie Academic and Professional: New York, 1995; Chapter 7. (c) Wayne, R. P. *Chemistry of Atmospheres*; Clarendon Press: Oxford, UK, 1996; pp 252−263.

(18) Bohn, B.; Zetzsch, C. *Phys. Chem. Chem. Phys.* **1999**, *1*, 5097.

(19) Bohn, B. *J. Phys. Chem. A* **2001**, *105*, 6092.

(20) Johnson, S.; Raoult, S.; Rayez, M. T.; Rayez, J. C.; Lesclaux, R. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4678.

(21) Grebenkin, S. Y.; Krasnoperov, L. N. *J. Phys. Chem. A* **2004**, *108*, 1953.

(22) Raoult, S.; Rayez, M. T.; Rayez, J. C.; Lesclaux, R. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2245.

(23) Ghigo, G.; Tonachini, G. *J. Am. Chem. Soc.* **1999**, *121*, 8366.

(24) Tuazon, E. C.; MacLeod, H.; Atkinson, R.; Carter, W. P. L. *Environ. Sci. Technol.* **1986**, *20*, 383.

(25) Volkamer, R.; Platt, U.; Wirtz, K. *J. Phys. Chem. A* **2001**, *105*, 7865.

(26) Ghigo, G.; Tonachini, G. *J. Am. Chem. Soc.* **1998**, *120*, 6753.

(27) Motta, F.; Ghigo, G.; Tonachini, G. *J. Phys. Chem. A* **2002**, *106*, 4411.

(28) Marcus, R. A. *J. Phys. Chem.* **1968**, *72*, 891.

(29) Rayez, M. T.; Rayez, J. C.; Sawerysyn, J. P. *J. Phys. Chem.* **1994**, *98*, 11342.

(30) Berndt, T.; Böge, O. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4946.

(31) Volkamer, R.; Klotz, B.; Barnes, I.; Imamura, T.; Wirtz, K.; Washida, N.; Becker, K. H.; Platt, U. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1598.

(32) Berndt, T.; Böge, O. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1205.

(33) (a) Atkinson, R.; Carter, W. P. L.; Darnell, K. R.; Winer, A. M.; Pitts, J. N. *Int. J. Chem. Kinet.* **1980**, *12*, 779. (b) Atkinson, R.; Lloyd, A. C. *J. Phys. Chem. Ref. Data* **1984**, *13*, 315.

(34) (a) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214. (b) Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1.

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

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

(37) Stevens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(38) Hariharan, C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(39) (a) Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363. (b) Ishida, K.; Morokuma, K.; Kormornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153.

(40) (a) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154. (b) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

(41) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; John Wiley: New York, 1986; pp 86−87.

(42) For a review, see: Roos, B. O. *Adv. Chem. Phys.* **1987**, *69*, 399.

(43) See, for example: Zhao, J.; Zhang, R. *Adv. Quantum Chem.* **2008**, *55*, 177–213.

(44) For a review, see: Bartlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1967.

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

(46) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(47) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219.

(48) (a) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910. (b) Hampel, C.; Peterson, K. A.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1. (c) Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321.

(49) Jayatilaka, D.; Lee, T. J. *J. Chem. Phys.* **1993**, *98*, 9734.

(50) Rienstra-Kiracofe, J. C.; Allen, W. D.; Schaefer, H. F., III. *J. Phys. Chem. A* **2000**, *104*, 9823.

(51) (a) Anderson, K.; Malmqvist, P.-A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483. (b) Anderson, K.; Malmqvist, P.-A.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.

(52) McQuarrie, D. *Statistical Mechanics*; Harper and Row: New York, 1986.

(53) 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.

(54) Werner, H.-J.; Knowles, P. J.; Almlöf, J.; Amos, R. D.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn. A. J.; Eckert, S. T.; Hampel, C.; Leininger, C.; Lindh, R.; Lloyd, A. W.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Peterson, K. A.; Pitzer, R.; Pulay, P.; Rauhaut, G.; Schütz, M.; Stoll, H.; Stone, A. J.; Thorsteinsson, T. *MOLPRO, version 98.1*; University of Stuttgart: Germany, 1998.

(55) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(56) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfen-

nig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(57) Eyring, H. *J. Chem. Phys.* **1935**, *107*, 107.

(58) Truong, T. N.; Truhlar, D. G. *J. Chem. Phys.* **1990**, *93*, 1761.

(59) Eckart, C. *Phys. Rev.* **1930**, *35*, 1303.

(60) Benson, S. W. *Thermochemical Kinetics*; Wiley & Sons: New York, 1976; p 8.

(61) (a) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon: Oxford, UK, 1990. (b) Matta, C. F.; Boyd, R. J. An Introduction to the Quantum Theory of Atoms in Molecules. In *The Quantum Theory of Atoms in Molecules*, 1st ed.; Matta, C. F., Boyd, R. J., Eds.; Wiley−VCH: Weinheim, Germany, 2007; pp 1−34.

(62) Suh, I.; Zhang, R.; Molina, L. T.; Molina, M. J. *J. Am. Chem. Soc.* **2003**, *125*, 12655.

(63) Fan, J.; Zhang, R. *J. Phys. Chem. A* **2006**, *110*, 7728.

(64) Fan, J.; Zhang, R. *J. Phys. Chem. A* **2008**, *112*, 4314.

CT900082G

JCTC Journal of Chemical Theory and Computation

Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules

Daniel J. Sindhikara,[†] Seonah Kim,[§] Arthur F. Voter,[||] and Adrian E. Roitberg*,[‡]

Quantum Theory Project and Departments of Physics and Chemistry, University of Florida, Gainesville, Florida 32611, Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, and Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Received December 23, 2008

Abstract: Molecular dynamics simulations starting from different initial conditions are commonly used to mimic the behavior of an experimental ensemble. We show in this article that when a Langevin thermostat is used to maintain constant temperature during such simulations, extreme care must be taken when choosing the random number seeds to prevent statistical correlation among the MD trajectories. While recent studies have shown that stochastically thermostatted trajectories evolving within a single potential basin with identical random number seeds tend to synchronize, we show that there is a synchronization effect even for complex, biologically relevant systems. We demonstrate this effect in simulations of alanine trimer and pentamer and in a simulation of a temperature-jump experiment for peptide folding of a 14-residue peptide. Even in replica-exchange simulations, in which the trajectories are at different temperatures, we find partial synchronization occurring when the same random number seed is employed. We explain this by extending the recent derivation of the synchronization effect for two trajectories in a harmonic well to the case in which the trajectories are at two different temperatures. Our results suggest several ways in which mishandling selection of a pseudorandom number generator initial seed can lead to corruption of simulation data. Simulators can fall into this trap in simple situations such as neglecting to specifically indicate different random seeds in either parallel or sequential restart simulations, utilizing a simulation package with a weak pseudorandom number generator, or using an advanced simulation algorithm that has not been programmed to distribute initial seeds.

1. Introduction

The use of molecular-dynamics (MD) simulations is widespread across various fields.[1] It is often useful to perform MD simulations in the canonical ensemble (NVT) in order to compare with experimental processes. In such circumstances, a thermostat is used to regulate temperature. Many types of thermostats are commonly employed, including Berendsen,[2] Nose-Hoover,[3–5] Andersen,[6] and Langevin.[7] Andersen and Langevin are stochastic in nature -- including random forces to mimic the effect of solvent collisions. Both can be proven[8,9] to give true canonical sampling. The characteristics of simulations using these stochastic thermostats, especially the commonly used Langevin thermostat, are the primary focus of this manuscript. An excellent review of characteristics of various thermostats can be found in ref 9.

Biomolecular simulations run at constant energy or constant temperature using a nonstochastic thermostat, such

* Corresponding author e-mail: roitberg@qtp.ufl.edu.

[†] Quantum Theory Project and Department of Physics, University of Florida.

[‡] Quantum Theory Project and Department of Chemistry, University of Florida.

[§] University of California.

[||] Los Alamos National Laboratory.

10.1021/ct800573m CCC: $40.75
Published on Web 04/28/2009

as Berendsen or Nose-Hoover, are considered to be chaotic and thus extremely sensitive to initial conditions. Braxenthaler et al. found that for a peptide system, root-mean-square deviations between two simulations can grow from only 0.001 Å to roughly 1 Å after only one or two picoseconds.[10] One might then expect that a stochastic thermostat, due to its use of random forces, would increase this divergent behavior. We will show in this article that under certain conditions this assumption is untrue, and failure to recognize this can lead to incorrect simulation results.

Stochastic thermostats use sequences of pseudorandom numbers to mimic the random solvent impacts. Pseudorandom number generators (PRNGs) are deterministic; given an initial 'seed', they always produce the same sequence of numbers. This trait is useful in that it allows for reproducibility of results when needed. Thus, if multiple simulations are run with different initial conditions, $(\vec{x},\vec{v})$, but identical random seeds, their random forces will remain the same for all simulations for their full length. Uberuaga et al. recently showed[11] that for the case of dynamics in a simple harmonic potential basin, Langevin (or Andersen) trajectories with identical seeds are *driven* to synchronize — the difference in both $\vec{x}$ and $\vec{v}$ between two trajectories decay exponentially, ultimately leading to a single trajectory path, no matter how different the initial conditions were. More generally, they argued that for a convex potential basin or even more general confining potentials a similar synchronization effect should occur. This behavior is consistent with rigorous mathematical results stating that under fairly general conditions, when the largest Lyapunov exponent is negative, trajectories starting from any ensemble of initial conditions are attracted to "random sinks".[12] Other groups have also observed this synchronization effect in model systems.[13–17] Cerutti et al. recently described[18] a situation where rapid restarts with the same seed of a single MD simulation in Langevin or Andersen Dynamics would result in a residual (nonzero average) stochastic force. We note, though, that this residual force is not the same as the synchronization effect.

In this paper, we show that synchronization can also occur in the much more complex potential energy landscapes of biomolecular systems. The potential energy surfaces for these systems typically consist of a complex network of many local minima separated by negatively curved saddle regions. Nonetheless, we observe that use of the same random number seed for different trajectories leads to strongly biased behavior due to partial synchronization occurring on the typical simulation time scale. We show this for Langevin dynamics of small peptides (alanine trimer and pentamer) and a simulation of a temperature-jump experiment for peptide folding of a 14-residue peptide.

We also explore the possibility that trajectories with identical seeds at different temperatures can synchronize, extending the harmonic-well derivation of Uberuaga et al. to the case in which the two trajectories have different temperatures. It will be shown that there is a well-defined synchronization of the coordinates, and hence a strong correlation between the trajectories exists. This multiple temperature synchronization has important implications for the method of replica exchange molecular dynamics among temperatures (T-REMD)[19,20] and variants thereof.[21–23] T-REMD is an enhanced sampling algorithm commonly used in the biomolecular simulation community. If the same random number seed is used for all the replicas, correlations among the different trajectories will contaminate the statistics of the study.

We first review the derivation of the driven synchronization for a pair of trajectories in a harmonic oscillator and then extend it to the case of two trajectories at different temperatures. We then present results from various peptide simulations in which the synchronization effect causes a bias in the results, culminating with the case of the T-REMD simulations of alanine trimer. We close with a discussion of the importance of understanding, and avoiding, the statistical contamination that can be caused by this synchronization effect in biomolecular simulations, and we identify and explain several common situations in which a simulator may unknowingly initiate multiple trajectories with the same random number seed, including neglecting to distribute random seeds for simultaneous simulations or sequential restart simulations or using programs that do not enforce distributions of random seeds.

2. Theory

2.1. Single Temperature Langevin Synchronization. In Langevin dynamics, particles are propagated based on the Langevin equation of motion:

$$m_i\ddot{\vec{r}}_i = -\vec{\nabla}V(\vec{r}_i) - \gamma m_i\dot{\vec{r}}_i + \vec{A}(\gamma, T, \upsilon) \qquad (1)$$

Here m_i, $\ddot{\vec{r}}_i$, $\dot{\vec{r}}_i$, and $\vec{r}_i$ are the mass, acceleration, velocity, and position of the i^{th} particle, respectively. $V(\vec{r}_i)$ is the potential energy determined by the force field. Equation 1 is essentially Newton's second law with two extra terms: a solvent drag force represented by $\gamma m_i\dot{\vec{r}}_i$ and a random force, $\vec{A}$, which obeys the fluctuation−dissipation theorem: $\langle A_i(t)A_j(t+\Delta t)\rangle = 2m\gamma k_bT\delta(\Delta t)\delta_{ij}$. Here the average, $\langle\rangle$, is over time, k_b is the Boltzmann constant, and $\delta(\Delta t)$ and δ_{ij} represent the Dirac and Kronecker delta functions, respectively. The magnitude and direction of A are based on a pseudorandom number, υ, and a probability distribution based on the temperature and heat bath coupling strength, γ, also known as the collision frequency, or friction. It has been shown[11] that for two particles in the same harmonic well with the same random number sequence, their trajectories are driven to synchronize. That is, for the i^{th} degree of freedom in a single dimension, $\Delta x_i = x_i^a - x_i^b$, the difference between two trajectories a and b, tends to zero as time increases.

Let us first consider the differences (between trajectories a and b) in the instantaneous accelerations on each degree of freedom in the Langevin regime:

$$\Delta\ddot{x} = -(\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma\Delta\dot{x} + (A_a - A_b)/m \qquad (2)$$

where $\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b$ and $\Delta\dot{x} = \dot{x}_a - \dot{x}_b$. If we approximate the local potential region to be a harmonic well of index a or b, $\partial V/\partial x = m\omega_{a,b}^2x$, then the difference in accelerations becomes $\Delta\ddot{x} = -\omega_a^2(x_a - x_b\omega_b^2/\omega_a^2) - \gamma\Delta\dot{x} + (A_a(t) - A_b(t))/m$. If the same pseudorandom number initial seed is used

for both simulations, the difference in random forces becomes zero at every step. What is then left is $\Delta\ddot{x} = -\omega_a^2(x_a - x_b\omega_b^2/\omega_a^2) - \gamma\Delta\dot{x}$. If the basins have identical curvature (or if the two simulations are in the same basin), this reduces to $\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x}$, which is the equation of a damped harmonic oscillator. For long times, the difference in the coordinates becomes zero; i.e. the trajectories 'synchronize'.

Realistic systems are more complicated than a simple harmonic oscillator. For these systems, synchronization rates are disrupted by a passage of particles through regions of negative curvature.[11] Despite this, any bound system must inevitably exist in some greater basin. Thus, synchronization may eventually occur for almost any simulated system unless a different initial seed for the PRNG is used.

Even before complete synchronization occurs for these many-minima systems, partial synchronization may occur for particles in basins of similar shape (this is where the shift from the identical harmonic well solution is small). As we will show, partial synchronization between trajectories does indeed take place on the time-scale of realistic simulations of peptide systems and has an effect strong enough to corrupt the results.

2.2. Multiple Temperature Langevin Synchronization. Though constant temperature simulations are both useful and commonplace, it is often necessary to use advanced simulation algorithms that utilize multiple temperatures for better sampling. One such enhanced sampling method that employs multiple-temperature simulation is parallel tempering[20] (PT), also known as replica exchange molecular dynamics among temperatures[19] (T-REMD). In this approach, replicas of the same molecule are simulated in parallel at different temperatures, most of which are above physiological temperatures. Periodically, a Metropolis-style Monte Carlo swap is attempted between conformations at different temperatures. A discussion of many generalized ensemble algorithms for enhanced sampling including T-REMD can be found in a review by Okamoto.[24]

The analytical derivation shown in the previous section suggests that single temperature simulations should synchronize. One might expect that the added complication of multiple temperatures might diminish synchronization. We show that even when using multiple temperatures, this effect is present and of consequence. We follow the same scheme and notation as we did in the single temperature derivation with two trajectories a and b, employing the same sequence of random numbers, but at two temperatures, T_a and T_b. Following the fluctuation dissipation theorem, we see that the stochastic forces are related by a simple scaling factor, $A_i^b(t) = cA_i^a(t)$, where $c = \sqrt{T_b/T_a}$. The equation for the difference in the i^{th} degree of freedom between the two runs is then given by

$$\Delta\ddot{x} = -(\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma\Delta\dot{x} + (1 - c)A_a/m \tag{3}$$

where $\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b$. As before, for a single harmonic potential with one degree of freedom, $\partial V/\partial x = m\omega^2 x$, and this equation becomes

$$\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x} + (1 - c)A_a/m \tag{4}$$

When the two temperatures are the same, $c = 1$, and this simplifies to the damped harmonic oscillator equation. When the temperatures differ by a small amount, c is close to unity, and the equation of motion for the difference between the two trajectories, Δx, becomes a Langevin equation with only a small noise term.

We can be more specific about how the two trajectories differ for any two temperatures by using a simple rescaling argument. Equation 3 can be modified to elucidate this behavior:

$$c\ddot{x}_a - \ddot{x}_b = -(c\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma(c\dot{x}_a - \dot{x}_b) + (c - c)A_a/m \tag{5}$$

The last term (the noise) vanishes, and, for a harmonic oscillator, the linearity allows us to simplify this to

$$c\ddot{x}_a - \ddot{x}_b = -\omega^2(cx_a - x_b) - \gamma(c\dot{x}_a - \dot{x}_b) \tag{6}$$

Defining $y = cx_a - x_b$, this becomes $\ddot{y} = -\omega^2 y - \gamma\dot{y}$, which is again simply a damped harmonic oscillator. Thus the trajectory for cx_a synchronizes to the trajectory for x_b; i.e. the trajectories at the two temperatures are related by a simple rescaling by c. Knowing the trajectory at any temperature is sufficient to specify exactly what the trajectory at any other temperature will be, once they have run long enough to be synchronized. Since the trajectories are now strongly correlated, the effective sampling will be greatly diminished.

3. Methods

3.1. Single Temperature Simulations. Single temperature MD simulations were performed on three peptides: trialanine (Ac-AAA-NH_2, ALA_3), penta-alanine (Ac-AAAAA-NH_2, ALA_5), and a 14-residue peptide (Ac-YGSPEAAA-KAAAA-r-NH_2, where r represents D-Arg). All simulations were performed using the AMBER 9 molecular simulation suite[25] with Langevin Dynamics in generalized-Born implicit solvent. All simulations were performed in AMBER 9 with the AMBER ff99SB force field,[26] and the Generalized Born implicit solvent model $GB^{(OBC)}$ was used to model the water environment in all our calculations.[27] The SHAKE algorithm[28] was used to constrain the bonds connecting hydrogen and heavy atoms in all the simulations. For the polyalanine peptides, a 1 fs integration time step was used, and each calculation was performed in the canonical ensemble (NVT) with a Langevin thermostat, using collision frequencies, γ, of 1 ps^{-1} or 50 ps^{-1} (as specified). For the 14-residue peptide, a 2 fs time step was used with a Langevin collision frequency of 5 ps^{-1}. For the 14-residue peptide, 1200 initial coordinate sets were taken from previously equilibrated run for the DS and SS production runs, which were run at an increased temperature of 372 K in order to simulate a Temperature-jump (T-jump) experiment.

To demonstrate single temperature trajectory synchronization, multiple simulations were run, all with different initial coordinates, using either the same initial random seed (SS) or different seeds (DS). 100 simulations each were run for 1 ns for polyalanines, 1200 simulations for 5 ns each for the 14-residue peptide.

3.2. Multiple Temperature Simulations. Synchronization across multiple temperatures is demonstrated by use of

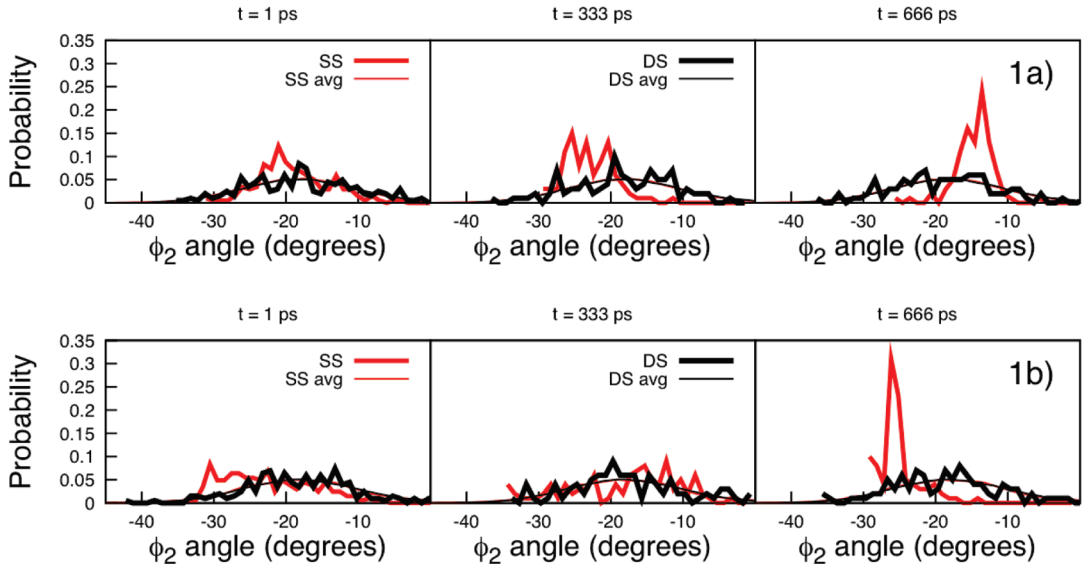


Figure 1. **a,b.** Probability distributions of the dihedral angle φ_2 across sets of 100 simulations for alanine polymer simulations. Data from SS simulations are shown in red, DS in black. Parts Figures a and b show ALA3 with $\gamma = 1$ ps^{-1} and $\gamma = 50$ ps^{-1}, respectively.

T-REMD simulation. Both a DS and SS 100-ns T-REMD simulations were performed using the AMBER9 package with Langevin dynamics in implicit solvent GB model. A Langevin thermostat was used with a collision frequency of 50 ps^{-1}. The SHAKE algorithm was employed allowing use of a 2 fs time step. Both systems utilized 6 replicas and started from the same initial configurations. The replica temperatures were spaced geometrically: 251.8 K, 300.0 K, 357.5 K, 426.0 K, 507.6 K, and 604.8 K. Exchanges were attempted every 500 steps (1 ps). The T-REMD code was altered to keep the random number sequences synchronized for all replicas for the SS simulation. Snapshots were recorded every 25 ps.

4. Results and Discussion

4.1. Single Temperature Synchronization. For both the ALA_3 and ALA_5 simulations, the dihedral angle of the second residue (ϕ_2) was measured versus time as an internal unit. We could have chosen any other set of coordinates to illustrate the synchronization effect.

Figure 1a,b shows probability distributions of ϕ_2 across the sets of 100 simulations for ALA_3 for a collision frequency of 1 ps^{-1} or 50 ps^{-1}, respectively. According to the harmonic theory,[11] in the low collision frequency regime, increasing the frequency, γ, should yield faster synchronization of trajectories. Histograms are shown at arbitrary intervals of 1 ps, 333 ps, and 666 ps into the trajectory. For comparison, probability distributions across the entire trajectories are shown in thin lines though since nearly identical, they are virtually indistinguishable.

Regardless of the random seeds (SS or DS), after a very long time, the distributions of ϕ_2 angles (very thin lines) are the same in each case. If synchronization is not present (as in the case of DS), the distribution of ϕ_2 among the 100 trajectories at a given time should be similar to the longer time average population. Conversely, if the same seeds (SS) are used for all 100 trajectories, the system behaves very differently. For instance, at 666 ps, for parts a and b of Figure 1, a large number of trajectories have very similar values of ϕ_2, as represented by a sharply peaked histogram. Figure 1 clearly shows that even in complex systems, the effect of synchronization is observable. The behavior of a coordinate for a set of SS simulations is similar to a swarm that expands and tightens as if compelled to come together. A movie of the Ramachandran plot (in beta/ppII region) of the first residue of ALA_3 with $\gamma = 50$ ps^{-1} demonstrates this behavior (Supporting Information).

To quantify synchronization among the entire set of 100 simulations, we used a measure of how many of them were similar to each other at any particular time. This was done by histogramming a physical observable (again ϕ_2 in our case) and counting how many of the 100 simulations reside in the histogram bin with maximum population. This is equivalent to the maximum height in Figure 1. If all 100 systems were perfectly synchronized, the highest fractional population (HFP) would be exactly one. Conversely, for completely unsynchronized systems, the HFP should stay relatively constant (and small for small bin sizes). For our simulations, the frame-by-frame ϕ_2 population was binned in narrow 2-degree windows. Figure 2a,b displays the time series of the fractional population of the most popular bin (HFP) for both ALA_3 and ALA_5, versus time. The sideplots show the probability distributions of those HFP time series. Included in the SI are additional HFP time series and probability distributions.

From the figures it is clear that the DS simulations have a small and relatively constant HFP. This is expected since the trajectories evolve independently from each other (there are not many simulations where the ϕ_2 angles are the same). In contrast, the SS simulations HFPs are much larger than for the DS case and in some cases achieve extremely high values. For instance, at 628 ps, a HFP value of 0.78 (Figure 2a) means that 78 of the 100 simulations have the same value of ϕ_2 (to within 2 degrees). The HFP difference between

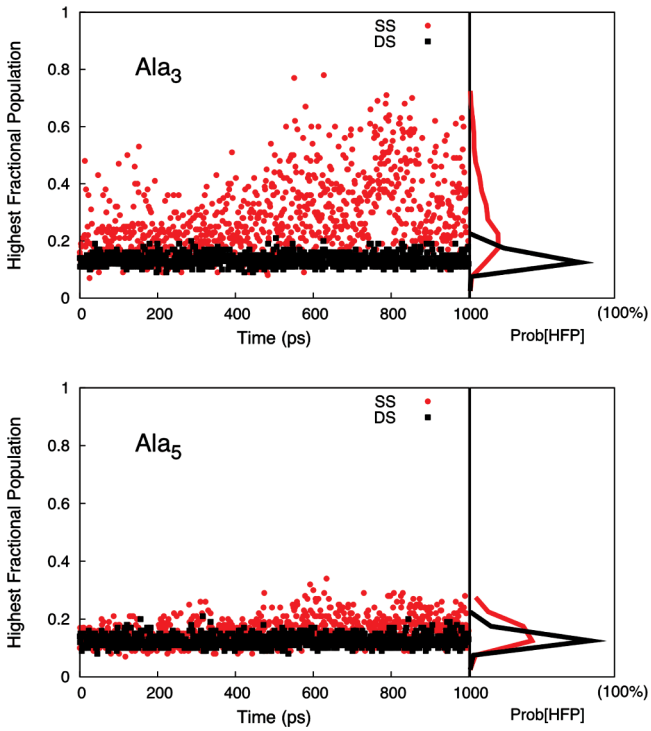


Figure 2. **a,b.** Highest fractional population (HFP) for φ_2 in alanine polymer simulations with a collision frequency, γ, of 50 ps^{-1}. Parts a and b show HFP for ALA_3 and ALA_5, respectively. Sideplots are the appropriate probability distribution of HFPs.

DS and SS is significantly smaller for the ALA_5 possibly because the synchronization is slower for larger systems. However, large homogeneous systems, such as those with explicit solvent, may still synchronize quickly.

To visualize the system as a whole, a time snapshot from the ALA_3 simulations was chosen and shown with all 100 simulation frames superimposed (see Figure 3, created in VMD[29]). In this figure, the red and blue spheres represent atom locations in the SS and DS simulations respectively. A stick representation is shown in gray as a visual aid. There is the same number of red spheres as blue (100 per atom). The figure shows fluctuation among the DS snapshots is much greater than that for SS; since the SS simulations are partially synchronized, the atomic positions are more condensed than they should be otherwise.

When simulating a complex system, it is often useful to utilize many simulations to reduce the error. The average over many simulations of a property, A, at time t,$\langle A\rangle_{sim}(t)$, is likely to be closer to the true average, $\bar{A}$, than the value of a property of a single simulation, $A(t)$ since value of A will fluctuate naturally in time. If the simulations are uncorrelated with each other, it can be shown that the standard deviation over time of these averages over simulations, $\sigma_{time}(\langle A\rangle_{sim})$, is less than the standard deviation over time of a single simulation $\sigma_{time}(A)$ (which is caused by the natural fluctuations):

$$\sigma_{time}(\langle A\rangle_{sim}) = \sigma_{time}(A)/\sqrt{N_{sim}} \tag{7}$$

Here, N_{sim} is the number of simulations. However, if the simulations are correlated with each other, the average over

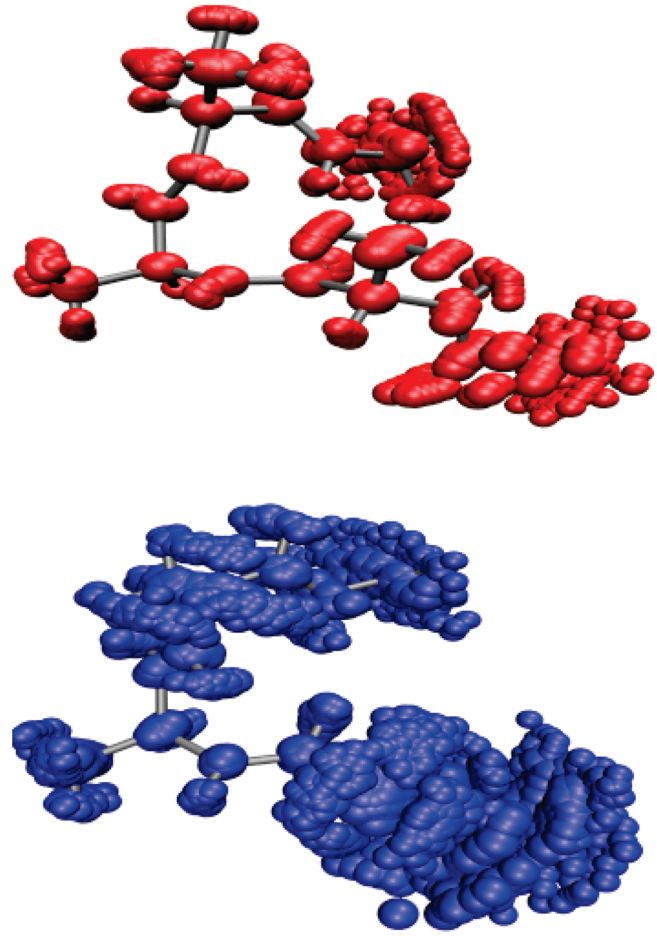


Figure 3. Sphere representations of simultaneous frames of 100 simulations at 836th ps of ALA_3. Red spheres represent atoms in SS simulations; blue spheres represent atoms in DS simulation.

simulations $\langle A\rangle_{sim}$ will fluctuate in time with greater amplitude, similar to that of a single simulation, $A(t)$. This has the same effect as reducing N_{sim}. Thus, according to eq 7, correlated simulations will have a higher standard deviation over time of the average over simulations, $\sigma_{time}(\langle A\rangle_{sim})$. We have presented above some arguments and results showing that many simulations run with the same initial random number generator seed will become somewhat synchronized over time. This effect will cause correlations between different simulations.[30]

We present here a striking example of this effect demonstrated in a simulation of temperature jump folding for a 14-residue peptide. This peptide was chosen since the T-jump kinetics were recently measured experimentally.[31] We have previously published a protocol for the simulation of that type of experiment.[32] In physical T-jump experiments, proteins are heated rapidly by a laser to observe folding events. Typically, a spectroscopic measure such as Trp-fluorescence or IR absorbance is used to follow the subsequent population relaxation. Unfortunately, in simulations, these phenomena are difficult to estimate. The expected CD spectra, rather, can be estimated in simulations based on the structure of the system. Since the CD signal at 222 nm is

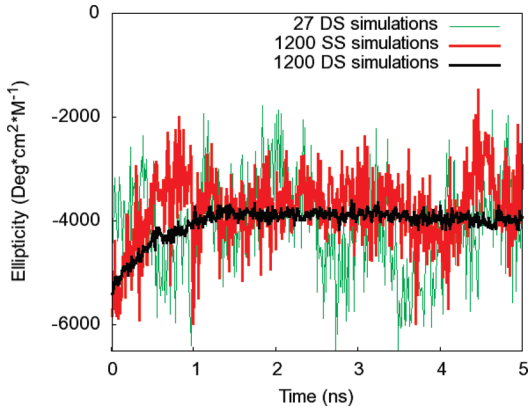


Figure 4. Average ellipticity vs time for a simulated T-jump experiment averaging over 1200 trajectories. SS simulations are shown in red, DS in black. The average of only 27 DS simulations is shown in green.

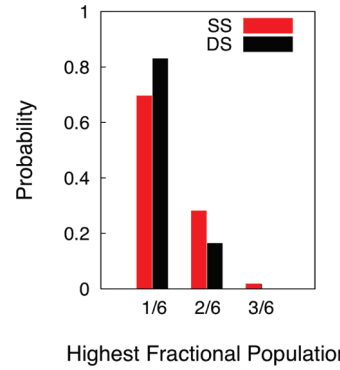


Figure 5. Histogram of highest fractional population for φ_2 angle in 2-degree bins for ALA_3 T-REMD simulation. SS shown in red, DS shown in black.

sometimes used to measure average ellipticity of the molecules, we focused on this measure to observe T-jump kinetics.

We computed ellipticity at 222 nm vs time averaged over 1200 simulations, using the method introduced by Sreerama and Woody.[33] Figure 4 shows the ellipticity vs time averaged over all 1200 simulations for SS (red line) and DS (black line), respectively. As can be seen in the figure, the signal-to-noise ratio is dramatically worse for the SS simulations.

The standard deviation for the last 2.5 ns of the T-jump simulation is 689 and 104 deg*cm^2*dM^{-1} for the SS and DS simulations, respectively. Thus, according to eq 7, the effective number of simulations is 44 ($\sqrt{689/104}$) times smaller for the SS than the DS. According to our preceding explanation, this means that the single seed runs act not like 1200 runs but as if only 27 (1200/44) truly independent runs. Thus, the average over 27 DS simulations (1200/44) should have a similar standard deviation to that of the 1200 SS simulations (thin green line in Figure 4). This effect is clearly shown in Figure 4 as a thin green line. We clarify that the single seed runs are not 'wrong' but that they produce overall fluctuations that are equivalent to a much smaller number of independent runs.

4.2. Multiple Temperature Synchronization. In the T-REMD simulations of alanine trimer, only six simultaneous simulations could be compared — one for each replica (as opposed to the 100 or 1200 simulations from the single temperature simulations above). Figure 5 shows the histogram of the highest fractional population of ϕ_2 bins (2 degree bins) for SS and DS simulations. The DS simulations (black bars) have a higher probability to have an HFP of 1/6, that is, that no two replicas have a ϕ_2 angle within 2 degrees of each other. The SS simulations are more likely to have two or three replicas with the same ϕ_2 angle. This indicates that some synchronization does occur between replicas. Not only such a simulation would be biased but also additional consequences for exchange probabilities may exist. Although T-REMD was used as an example, we expect synchronization to occur for any set of multiple temperature simulations.

4.3. Relevance of Synchronization. As evidenced by our results, thermostat induced trajectory synchronization biases results and should be avoided. Depending on the severity of the synchronization, the bias may or may not be obvious to the researcher. It is thus important to understand the nature of synchronization to be aware of situations where it might occur.

Synchronization occurs when there is an overlap of pseudorandom number sequences, and this is typically caused by using the same initial seed for multiple runs. This can happen inadvertently for many reasons. Some simulation programs use a default random seed. AMBER, for example, uses a default random seed if none is specified. Others may use a time-seeded PRNG, which, depending on the implementation, may give a high risk of giving identical seeds. For example, if the program uses a time seed connected with a clock that is discretized to the nearest second, then if many simulations are initiated simultaneously, there is a high probability that many or all will receive the same seed.

Quite often simulators restart simulations. If one restarts with the same parameters (including initial seed), then the simulations could become self-synchronized. Cerutti et al. recently reported[18] a different negative consequence of repeatedly restarting Langevin or Andersen MD runs with the same initial seed -- trajectory corruption caused by a nonzero average stochastic force. Additionally, coders of new methods, such as T-REMD, which string together MD segments, might unknowingly build a code that uses the same seed. As we have shown, even Langevin MD runs at different temperatures can become synchronized.

Furthermore, since PRNGs have an inherent period, MD runs which call the PRNG more than this amount will naturally repeat the sequence. Although advanced PRNGs such as the Marsaglia algorithm[34] have extremely long periods (2^{144} for AMBERs implementation), older PRNGs have much shorter periods. We highly recommend that simulators take note of the PRNG period of the program they are running.

5. Conclusion

We have shown that identical-noise synchronization effects, previously observed for relatively simple systems under the influence of a stochastic thermostat, can also occur in the much more complex systems typical of

biomolecular simulations. Even in the case of trajectories at different temperatures, harmonic analysis shows a special scaled synchronization will occur. We indeed found evidence of synchronization bias in a replica-exchange simulation. In a simulation study, this synchronization tendency, even if weak, will corrupt the statistical quality of the results and may even lead to incorrect conclusions about the qualitative behavior of the system. Using modern biomolecular simulation programs and methods, many ways exist in which one can inadvertently initiate trajectories with identical seeds. It is possible that many papers have already been published with data that are biased by synchronization. We advise that great care be taken to avoid this situation by meticulous preparation of seeds, and we suggest that authors may wish to state specifically whether different initial seeds have been used when their results are based on multiple trajectories with a stochastic thermostat.

Acknowledgment. The authors acknowledge the University of Florida High-Performance Computing Center for providing computational resources. Computational resources were also provided by Teragrid Grant No. TG-MCA05S010. Work at the University of Florida was funded by the National Science Foundation grant number CHE-0822-935. Work at Los Alamos National Laboratory (LANL) was supported by the United States Department of Energy (U.S. DOE) Office of Basic Energy Sciences, Materials Sciences and Engineering Division. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. DOE under Contract No. DE-AC52-06NA25396. The authors are grateful to Blas P. Uberuaga, Marian Anghel, Kevin Lin, and John Chodera for helpful discussions.

Supporting Information Available: Additional plots containing statistical data (HFP) and visualization (3 figures, 1 movie). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tuckerman, M. E.; Martyna, G. J. Understanding modern molecular dynamics: Techniques and applications. *J. Phys. Chem. B* **2000**, *104* (2), 159–178.

(2) Berendsen, H. J. C.; Potsma, J. P. M.; van Gunsteren, W. F.; DiNola, A. D.; Haak, J. R. Molecular Dynamics with Coupling to and External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(3) Nose, S. A Molecular-Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52* (2), 255–268.

(4) Hoover, W. G. Canonical Dynamics - Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(5) Evans, D. J.; Holian, B. L. The Nose-Hoover Thermostat. *J. Chem. Phys.* **1985**, *83* (8), 4069–4074.

(6) Andersen, H. C. Molecular-Dynamics Simulations at Constant Pressure and-or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393.

(7) Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **1973**, *9* (3), 215–220.

(8) Andersen, H. C. Molecular Dynamics at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

(9) Hunenberger, P. Thermostat algorithms for molecular dynamics simulations. In *Advanced Computer Simulation Approaches for Soft Matter Sciences I*; Springer: Berlin/Heidelberg, 2005; Vol. 173, pp 105−147.

(10) Braxenthaler, M.; Unger, R.; Auerbach, D.; Given, J. A.; Moult, J. Chaos in protein dynamics. *Proteins: Struct., Funct., Genet.* **1997**, *29* (4), 417–425.

(11) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374.

(12) Le Jan, Y. Equilibre statistique pour les produits de difféomorphismes aléatoires indépendants. *Annales de l'I.H.P. Probabilités et statistiques* **1987**, *23* (1), 111–120.

(13) Fahy, S.; Hamann, D. R. Transition from Chaotic to Non-chaotic Behavior in Randomly Driven Systems. *Phys. Rev. Lett.* **1992**, *69* (5), 761–764.

(14) Maritan, A.; Banavar, J. R. Chaos, Noise, and Synchronization. *Phys. Rev. Lett.* **1994**, *72* (10), 1451–1454.

(15) Lise, S.; Maritan, A.; Swift, M. R. Langevin equations coupled through correlated noises. *J. Phys. A: Math. Gen.* **1999**, *32* (28), 5251–5260.

(16) Ciesla, M.; Dias, S. P.; Longa, L.; Oliveira, F. A. Synchronization induced by Langevin dynamics. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2001**, *6306* (6), -.

(17) Longa, L.; Curado, E. M. F.; Oliveira, F. A. Roundoff-induced coalescence of chaotic trajectories. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **1996**, *54* (3), R2201−R2204.

(18) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* **2008**, *4*, 1669–1680.

(19) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141–151.

(20) Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140–150.

(21) Hagen, M.; Kim, B.; Liu, P.; Friesner, R. A.; Berne, B. J. Serial replica exchange. *J. Phys. Chem. B* **2007**, *111* (6), 1416–1423.

(22) Shen, H. J.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. Implementation of a serial Replica Exchange Method in a physics-based united-residue (UNRES) force field. *J. Chem. Theory Comput.* **2008**, *4* (8), 1386–1400.

(23) Rick, S. W. Replica exchange with dynamical scaling. *J. Chem. Phys.* **2007**, *126* (5), 054102.

(24) Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.* **2004**, *22* (5), 425–439.

(25) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(26) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone

parameters. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (3), 712–725.

(27) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (2), 383–394.

(28) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(29) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics Modell.* **1996**, *14* (1), 33–&.

(30) Friedberg, R.; Cameron, J. E. Test of Monte-Carlo Method - Fast Simulation of a Small Ising Lattice. *J. Chem. Phys.* **1970**, *52* (12), 6049–6058.

(31) Wang, T.; Du, D. G.; Gai, F. Helix-coil kinetics of two 14-residue peptides. *Chem. Phys. Lett.* **2003**, *370* (5−6), 842–848.

(32) Kim, S.; Roitberg, A. E. Simulating temperature jumps for protein folding. *J. Phys. Chem. B* **2008**, *112* (5), 1525−1532.

(33) Sreerama, N.; Woody, R. W., Computation and analysis of protein circular dichroism spectra. *Methods Enzymol* **2004**, *383*, 318−351.

(34) Marsaglia, G.; Narasimhan, B.; Zaman, A. A Random Number Generator for Pcs. *Comput. Phys. Commun.* **1990**, *60* (3), 345–349.

CT800573M

JCTC Journal of Chemical Theory and Computation

ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale

M. J. Harvey,*,† G. Giupponi,‡ and G. De Fabritiis§

Information and Communications Technologies, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom, Department de Fisica Fundamental, Universitat de Barcelona, Carrer Marti i Franques 1, 08028 Barcelona, Spain, and Computational Biochemistry and Biophysics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain

Received February 4, 2009

Abstract: The high arithmetic performance and intrinsic parallelism of recent graphical processing units (GPUs) can offer a technological edge for molecular dynamics simulations. ACEMD is a production-class biomolecular dynamics (MD) engine supporting CHARMM and AMBER force fields. Designed specifically for GPUs it is able to achieve supercomputing scale performance of 40 ns/day for all-atom protein systems with over 23 000 atoms. We provide a validation and performance evaluation of the code and run a microsecond-long trajectory for an all-atom molecular system in explicit TIP3P water on a single workstation computer equipped with just 3 GPUs. We believe that microsecond time scale molecular dynamics on cost-effective hardware will have important methodological and scientific implications.

I. Introduction

The simulation of mesoscopic scales (microseconds to milliseconds) of macromolecules continues to pose a challenge to modern computational biophysics. While the fundamental thermodynamic framework behind the simulation of macromolecules is well characterized, exploration of biological time scales remains beyond the computational capacity routinely available to many researchers. This has significantly inhibited the widespread use of molecular simulations for in silico modeling and prediction.[1]

Recently, there has been a renewed interest in the development of molecular dynamics simulation techniques. D. E. Shaw Research[2] has fostered several significant algorithmic improvements including midpoint[3] and neutral-territory methods[4] for the summation of nonbonded force calculations, a new molecular dynamics package called Desmond[5] and Anton,[2] a parallel machine for molecular dynamics simulations that uses specially designed hardware. Other parallel MD codes, such as Blue matter,[6] NAMD,[7] and Gromacs4,[8] have been designed to perform parallel MD simulations across multiple independent processors, but latency and bandwidth limitations in the interconnection network between processors reduces parallel scaling unless the size of the simulated system is increased with processor count. Furthermore, dedicated, highly parallel machines are usually expensive and not reservable for long periods of time due to cost constraints and allocation restrictions.

A further line of development of MD codes consists of using commodity high-performance accelerated processors.[1] This approach has become an active area of investigation, particularly in relation to the Sony−Toshiba−IBM Cell processor[9] and graphical processing units (GPUs). Recently, De Fabritiis[9] implemented an all-atom biomolecular simulation code, CellMD, targeted to the architecture of the Cell processor (contained within the Sony Playstation3) that reached a sustained performance of 30 Gflops with a speed up of 19 times compared to the single CPU version of the code. At the same time, a port of the Gromacs code for implicit solvent models[10] was developed and used by the Folding@home distributed computing project[11] on a distrib-

* To whom correspondence should be addressed. E-mail: m.j.harvey@imperial.ac.uk, gianni.defabritiis@upf.edu.
† Imperial College London.
‡ Universitat de Barcelona.
§ Universitat Pompeu Fabra.

10.1021/ct9000685 CCC: $40.75
Published on Web 05/21/2009

uted network of Playstation3s. Similarly, CellMD was used in the PS3GRID.net project[12] based on the BOINC platform[13] moving all-atom MD applications into a distributed computing infrastructure.

Pioneers in the use of GPUs for production molecular dynamics[11] had several limitations imposed by the restrictive, graphics-orientated OpenGL programming model[14] then available. In recent years, commodity GPUs have acquired nongraphical, general-purpose programmability and undergone a doubling of computational power every 12 months, compared to 18−24 months for traditional CPUs.[1] Of the devices currently available on the market, those produced by Nvidia offer the most mature programming environment, the so-called compute unified device architecture (CUDA),[15] and have been the focus of the majority of investigation in the computational science field.

Several groups have lately shown results for MD codes which utilize CUDA-capable GPUs. Stone et al.[16] demonstrated the GPU-accelerated computation of the electrostatic and van der Waals forces, reporting a 5.4 times speed up with respect to a conventional CPU. Meel et al.[17] described an implementation for simpler Lennard−Jones atoms which achieved a net speed up of up to 40 times over a conventional CPU. Unlike the former, the whole simulation is performed on the GPU. Recently, Phillips et al. reported experimental GPU acceleration of NAMD[18] yielding speed ups of up to 7 times over NAMD 2.6. Pande et al. announced OpenMM,[19] a library of GPU kernels for MD, derived from their work on Folding@Home.[11] Though very fast, the OpenMM kernels use direct summation of nonbonded terms, making its use suitable mainly for implicit solvent systems.

More widely in the field of computational chemistry, investigation has proceeded into GPU acceleration[20] of a variety of quantum chemical methods, including quantum Monte Carlo,[21] correlation,[22] and self-consistent field[23] methods.

In this work, we report on a molecular dynamics program called ACEMD which is optimized to run on Nvidia GPUs and which has been developed with the aim of advancing the frontier of molecular simulation toward the ability to routinely perform *microsecond-scale* simulations. ACEMD maximizes performance by running the whole computation on the GPU rather than offloading only selected computationally expensive parts. We developed ACEMD to implement all features of a typical MD simulation including those usually required for production simulations such as particle−mesh Ewald (PME)[24] calculation of long-range electrostatics, thermostatic control, and bond constraints. The default forcefield format used by ACEMD is CHARMM[25] and Amber[26] force fields. ACEMD also provides a scripting interface to control and program the molecular dynamics run to perform complex protocols like umbrella sampling, steered molecular dynamics and sheared boundary conditions, and metadynamics simulations.[27]

II. GPU Architecture

The G80 and subsequent G200 generations of Nvidia GPU architectures are designed for data-parallel computation, in which the same program code is executed in parallel on many

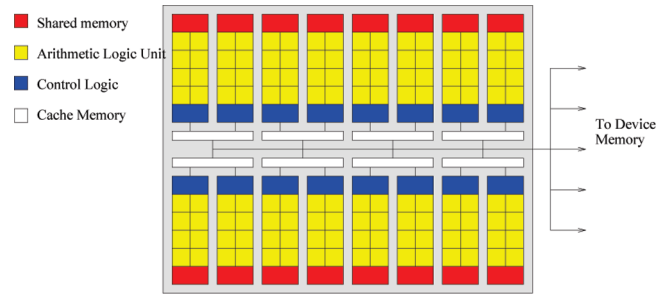


Figure 1. Nvidia GPU design is based around an 8-core single-program, multiple data (SPMD) processor. Each core has local storage provided by the register file and access to a shared memory region. Read/write access to the main device memory is uncached, except for some specific read-only access modes. The above figure represents the G80-series device with 16 such processors, while the contemporary G200 contains 30 (Tesla C1060, GTX280), giving 240 cores.

Table 1. Summary of Characteristics of First- and Second-Generation Nvidia GPU Compute Devices, with a Contemporary Intel Xeon Shown for Comparison[a]

	Tesla C870 (G80)	Tesla C1060 (G200)	Intel Xeon 5492
cores	128	240	4
clock (GHz)	1.350	1.296	3.4
mem bandwidth (MB/s)	77	102	21
Gflops (sp/dp)	512/−	933/78	108/54
power (W)	171	200	150
year	2007	2008	2008

[a] Data taken from manufacturers' data sheets. (sp stands for single precision and dp for double precision).

data elements. The CUDA programming model, an extended C-like language for GPUs, abstracts the implementation details of the GPU so that the programmer may easily write code that is portable between current and future GPUs. Nvidia GPU devices are implemented as a set of multiprocessor (MP) devices, each of which is capable of synchronously executing 32 program threads in parallel (called a warp) and managing up to 1024 concurrently (Figure 1).[55] Current Nvidia products based on these devices are able to achieve up to 933 Gflops in single precision. By comparison, a contemporary quad-core Intel Xeon CPU is capable of approximately 54 Gflops double precision/108 Gflops single precision.[28,29] A brief comparison of the characteristics of these devices is given in Table 1. Each MP has a set of 32-bit registers, which are allocated as required to individual threads, and a region of low-latency shared memory that is accessible to all threads running on it. The MP is able to perform random read/write access to external memory. Access to this global memory is uncached and so incurs the full cost of the memory latency (up to 400 cycles). However, when accessed via the GPU's texturing units, reads from arrays in global memory are cached, mitigating the impact of global memory access for certain read patterns. Furthermore, the texture units are capable of performing linear interpolation of values into multidimensional (up to 3D) arrays of floating point data.

While the older G80 architecture supported only single-precision IEEE-754 floating point arithmetic, the newer G200

design also supports double-precision arithmetic, albeit at a much lower relative speed. The MP has special hardware support for reciprocal square root, exponentiation, and trigonometric functions, allowing these to be computed with low latency but at the expense of slightly reduced precision.

Program fragments written to be executed on the GPU are known as *kernels* and executed in *blocks*. Each block consists of multiple instances of the kernel, called *threads*, which are run concurrently on a single multiprocessor. The number of threads in a block is limited by the resources available on the MP, but multiple blocks may be grouped together as a *grid*. The CUDA runtime, in conjunction with the GPU hardware itself, is responsible for efficiently scheduling the execution of a grid of blocks on available GPU hardware. CUDA does not presently provide a mechanism for transparently using multiple GPU devices for parallel computation. For full details of the CUDA environment, the reader is referred to the SDK documentation.[30]

III. Molecular Dynamics on the GPU

ACEMD implements all features of an MD simulation on a CUDA-compatible GPU device, including those usually required for production simulations in the NVT ensemble (i.e., bonded and nonbonded force term computation, velocity-Verlet integration, Langevin thermostatic control, smooth Ewald long-range electrostatics (PME),[24,31] and hydrogen bond constraints). Also implemented is the hydrogen mass repartitioning scheme described in ref 32 and used, for instance, in codes such as Gromacs, which allows an increased time step of up to 4 fs. The code does not presently contain a barostat, so simulations in the NPT ensemble are not possible. However, it is noted that with large molecular systems, changes in volume due to the pressure control are very limited after an initial equilibration making NVT simulations viable for production runs. ACEMD supports the CHARMM27 and Amber force fields, PDB, PSF, and DCD file formats,[33] check pointing, and input files compatible with widely used MD codes. It is extensible via a C-based plugin interface and TCL scripting (see ACEMD online manual for details[34]).

The computation of the nonbonded force terms dominates the computational cost of MD simulation, and it is therefore important to use an efficient algorithm. As in refs 9 and 17, we implement a cell-list scheme in which particles are binned according to their coordinates. On all-atom biomolecular systems a cutoff of $R = 12$ Å with bins $R/2$ gives an average cell population of approximately 22 atoms. This is comparable to the warp size of 32 for current Nvidia GPUs. In practice, however, transient density fluctuations can lead to the cell population exceeding the warp size. Consequently, the default behavior of ACEMD is to assume a maximum cell population of 64. The code may also accommodate a bin size of R for coarse-grained simulations. The cell-list construction kernel processes one particle per thread, with each thread computing the cell in which its atom resides. To permit concurrent manipulation of a cell-list array, atomic memory operations are used.

The nonbonded force computation kernel processes a single cell per thread block, computing the full Lennard−Jones and electrostatic force on each particle residing within it. All of the cells within R of the current cell (including a copy of the cell itself) are loaded into shared memory in turn. Each thread then computes the force on its particle by iterating over the array in shared memory. In contrast to CPU implementations, reciprocal forces are not stored for future use (i.e., the force term F_{ij} is not saved for reuse as F_{ji}), because of the relatively high cost of global memory access.[56] The texture units are used to assist the calculation of the electrostatic and van der Waals terms by providing linearly interpolated values for the radial components of those functions from lookup tables. The interpolation error is low and does not affect the energy conservation properties of NVE simulations (Figure 2).

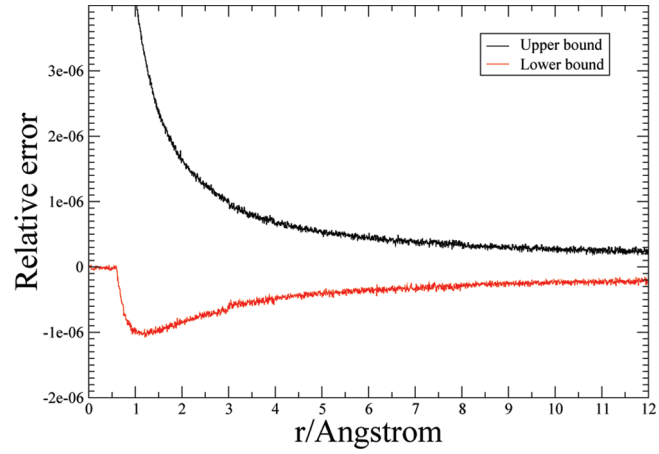


Figure 2. Bounds (running average) of the relative error $(E_{interp} - E_{calcd})/(E_{calcd})$ between directly calculated and lookup table (linear interpolation, $n = 4096$) values of the van der Waals potential. Error has a period of $(R_{max})/(n)$.

In production runs, the relative force error compared to a reference simulation performed in double precision is consistently less than 10^{-4}, below the 10^{-3} error considered the maximum acceptable for biomolecular simulations.[5] Particle-mesh Ewald (PME)[31] evaluation of long-range electrostatics is also supported by a dedicated kernel. All parts of this computation are performed on the GPU, with support from the Nvidia FFT library.[35] For PME calculations, a cutoff of $R = 9.0$ Å accepted as provided sufficient accuracy, permitting the maximum cell population to be limited to 32 atoms.

To support CHARMM and AMBER force fields, it is necessary to selectively exclude or scale nonbonded force terms between atoms that share an explicit bond term. The indices of excluded and 1−4 scaled pairs are stored in bitmaps, allowing any pair of particles with indices i,j such that $|i - j| \leq 64$ to be excluded or scaled.[16] Exclusions with larger index separations are also supported in order to accommodate, for example, disulfide bonds, but the additional book-keeping imposes a minor reduction in performance. Because the atoms participating in bonded terms are spatially localized, it is necessary only to make exclusion tests for interactions between adjacent cells despite, for cells of $R/2$, the interaction halo being two cells thick. A consequent optimization is the splitting of the nonbonded

Table 2. Energy Change in the NVE Ensemble per Nanosecond per Degree of Freedom (dof) in K_bT Units for Dihydrofolate Reductase (DHFR) Using Different Integration Time Steps, Constraints, and Hydrogen Mass Repartitioning (HMR) Schemes

time step (fs)	constraints	HMR	K_bT/ns/dof
1	no	no	0.00021
2	yes	no	−0.00082
4	yes	yes	−0.00026

force kernel into two versions, termed *inner* and *outer*, which, respectively, include and omit the test.

Holonomic bond constraints are implemented using the M-shake algorithm[36] and RATTLE for velocity constraints[37] within the velocity Verlet integration scheme.[38] M-shake is an iterative algorithm, and in order to achieve acceptable convergence it is necessary to use double-precision arithmetic (a capability available only on G200/architecture 1.3 class devices). For the pseudorandom number source for the Langevin thermostat we use a Mersenne twister kernel, modified from the example provided in the CUDA SDK.

IV. Single-Precision Floating-Point Arithmetic Validation

ACEMD uses single-floating point arithmetic because the performance of GPUs on single precision is much higher than double precision and the limitation of a single floating point can be controlled well for molecular dynamics.[8,9] Nevertheless, we validate in this section the conservation properties of energy in a NVT simulation using rigid and harmonic bonds, as constraints have shown to be more sensitive to numerical precision. Potential energies were checked against NAMD values for the initial configuration of a set of systems, including disulfide bonds, ionic system, protein, and membranes, in order to verify the correctness of the force calculations by assuring that energies were identical within 6 significant figures. The Langevin thermostat algorithm was tested for three different damping frequencies $\gamma = 0.1$, 0.5, and 1.0 with a reference temperature of $T = 300$ K and both with and without constraints.

The test simulations consist of nanosecond runs of dihydrofolate reductase (DHFR) joint AMBER-CHARMM benchmark with volume 62.233 × 62.233 × 62.233 Å^3 (a total of 23 558 atoms).[5] Each simulation system was first equilibrated at a temperature of T= 300 K and then relaxed in the NVE ensemble. A reference simulation with harmonic bonds and time step $dt = 1$ fs was also performed, as well as simulations with $dt = 2$ fs using rigid constraints and with $dt = 4$ fs with rigid constraints and hydrogen mass repartitioning (HMR) with a factor 4.0, as in ref 39. In Table 2 we show the energy change per nanosecond per degree of freedom in units of K_bT, which is similar to other single- and double-precision codes MD.[40] We note that even when using bigger time steps and a combination of M-shake and hydrogen mass repartitioning, energy conservation is reasonably good and much slower than the time scale at which the thermostat would act. Hydrogen mass repartitioning is an elegant way to increase the time step up to 4 fs by increasing the momentum of inertia of groups of atoms bonded to hydrogen atoms. The mass of the bonded heavy atoms to hydrogens is repartitioned among hydrogen atoms, leaving the total mass of the system unchanged. As individual atom masses do not appear in the expression for the equilibrium distribution, this repartition affects only the dynamic properties of the system not the equilibrium distribution. Following ref 32, a factor 4 for hydrogens affects only marginally the diffusion and viscosity of TIP3P water (which is in any case inaccurate when compared to experimental data). A similar speed up could also be obtained by using a smaller time step with the evaluation of the long-range electrostatic terms every other time step.

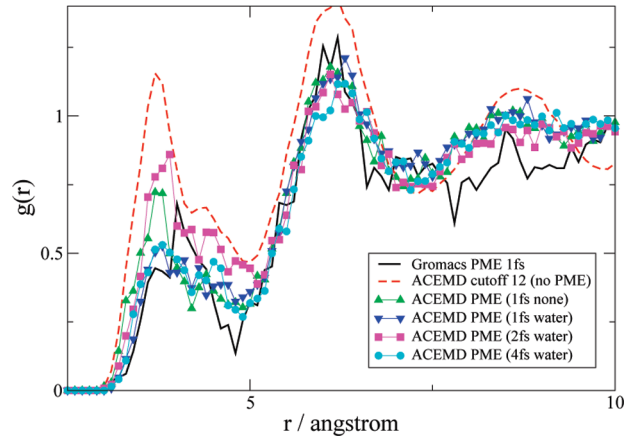


Figure 3. Plot of Na−Na pair distribution functions for a 1 M NaCl water box as in ref 41.

We also validated the implementation of the PME algorithm to compute long-range electrostatics forces. We ran a set of simulations using different time steps and algorithms as above ($dt = 1$ and 2 fs rigid bonds, $dt = 4$ fs rigid bonds and hydrogen mass repartitioning) on a 40.5 × 40.5 × 40.5 Å^3 box of 1 M solution of NaCl in water (6461 atoms), as in ref 41. PME calculations were performed with a 64 × 64 × 64 grid size. Two simulations of the same system were used as reference: one with Gromacs[8] with PME as in ref 41 and the other using ACEMD with an electrostatic cutoff of 12 Å without PME. We calculated the Na−Na pair distribution function $g(r)$ in Figure 3 in order to compare the simulation results for different simulations and methods, as from ref 41 Na−Na $g(r)$ results as the quantity more sensitive to different methods for electrostatics calculations. We note that for all integration time steps used, ACEMD agrees well with the reference simulation made with Gromacs. In addition, using PME gives consistently better results than using a 12 Å cutoff for this simple homogeneous system, as expected. A direct validation of the pair distribution function with the hydrogen mass repartitioning method is also shown in Figure 3, comparing the $g(r)$ for time step = 1, 2, and 4.

V. Performance

The current implementation of ACEMD is parallelized in a task parallel manner designed to scale across just 3 GPUs attached to a single host system. A simple force−decomposition scheme[42] is used, in which each GPU computes a subset of the force terms. These force terms are summed

Table 3. Performance of ACEMD on the DHFR Benchmark[a]

Program	CPU Cores and GPUs	ms/step
ACEMD	1 CPU, 1 GPU (240 cores)	17.55
ACEMD	3 CPU, 3 GPU (720 cores)	7.56
NAMD2.6	128 CPU (64 nodes)	9.7
NAMD2.6	256 CPU (128 nodes)	7.0
Desmond	32 CPU (16 nodes)	11.5
Desmond	64 CPU (32 nodes)	6.3
Gromacs4	20 CPU (5 nodes)	7

[a] GPUs are Nvidia GTX 280. NAMD, Desmond and Gromacs performances are indicative of the orders of magnitude speed up obtained with GPUs and ACEMD as they are all performed on different CPU systems (from refs 5 and 8 Gromacs figures interpolated from Figure 6 of ref 8). Desmond and Gromacs use SSE vector instructions and single precision floating-point numbers.

by the host processor and the total force matrix transferred back to each GPU which then performs integration of the whole system. ACEMD dynamically load balances the computation across the GPUs. This allows the simulation of heterogeneous molecular systems and also accommodates variation due to host system architecture (for example, different speed GPUs or GPU-host links). For simulations requiring PME, a heterogeneous task decomposition is used, with a subset of GPUs dedicated to PME computation.

The performance benchmark is based on the DHFR molecular system with a cutoff of $R = 9$ Å, switched at 7.5 Å, $dt = 4$ fs, PME for long-range electrostatic with $64 \times 64 \times 64$ grid size, and fourth-order interpolation, M-shake constraints for hydrogen bonds and hydrogen mass repartitioning. All simulations were run on a PC equipped with 4 Nvidia GPU GTX 280 cards at 1.3 GHz (just 3 GPUs used for these tests), a quad-core AMD Phenom processor (2.6 GHz), MSI board with AMD790 FX chipset, 4GB RAM running Fedora Core 9, CUDA toolkit 2.0, and the Nvidia graphics driver 177.73. Performance results reported in Table 3 indicate that ACEMD requires 17.55 ms per step with the DHFR system and 7.56 ms per step when run in parallel over the 3 GPUs. As expected by the simple task decomposition scheme, ACEMD achieves a parallel efficiency of 2.3 over 3 GPUs. Further device-to-device communication directives may substantially improve these results as they will enable the use of spatial-decomposition parallelization strategies, such as neutral territory (NT) schemes.[4] Comparing directly the maximum performance of ACEMD on the DHFR system with results of various MD programs from ref 5 we obtain a performance approaching that of 256 CPU cores using NAMD and 64 using Desmond on a cluster with fast interconnect. Using hydrogen mass repartitioning and a time step of 4 fs integration time step it is possible to simulate trajectories of over 45 ns per day with 3 GPUs and almost 20 ns per day with a single GPU. A highly optimized code such as Gromacs4 requires only 20 CPU cores to deliver similar performance to 3 GPUs on DHFR,[8] but the calculations are not identical as, for instance, there are several optimizations applied to water.

Representative performance data for ACEMD and the GPU-accelerated version of NAMD[18] for the apoA1 benchmark system (92 224 atoms) is given in Table 4. Differences between the simulation and hardware configurations prevent a direct comparison, but it is salient to note that because the enhanced NAMD retains the spatial decomposition parallelism it is able to scale across multiple GPU-equipped hosts, while ACEMD is designed for optimal performance on a small number of GPUs.

Table 4. Performance of ACEMD and NAMD on the apoA1 Benchmark[a]

program	CPU cores and GPUs	ms/step
ACEMD	1 CPU, 1 GPU (1 node)	73.4
ACEMD	3 CPU, 3 GPU (1 node)	32.5
NAMD	4 CPU, 4 GPU (1 node)	87
NAMD	16 CPU, 16 GPU (4 nodes)	27
NAMD	60 CPU (15 nodes)	44

[a] ACEMD run using Nvidia GTX 280 GPUs ($R = 9$ Å, PME every step), NAMD ($R = 12$ Å, PME every 4 steps) run with G80-series GPUs (approximately half as fast) NAMD performance data taken from ref 18.

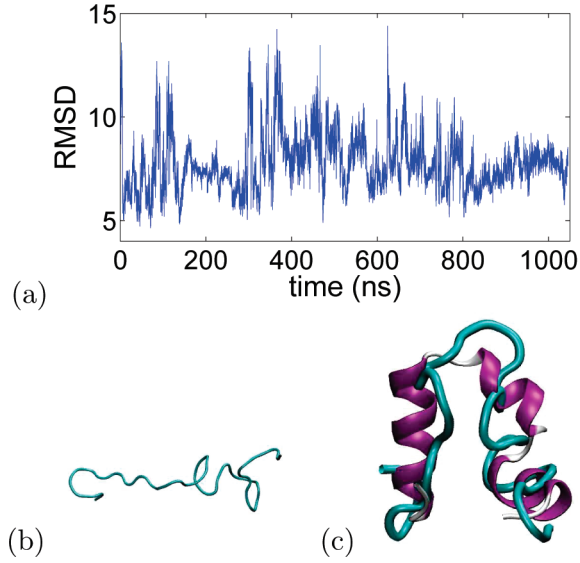


Figure 4. (a) rmsd of the backbone of the protein during the microsecond simulation starting from the unfolded configuration (b) of the Villin system with 13 701 atoms (TIP3P water not shown for clarity). Within our simulation window the minimum rmsd was 4.87 Å for which the resulting best structure is overlapped with the crystal structure in c.

VI. Microsecond Simulations on Workstation Hardware

To provide a direct demonstration that molecular simulations have now entered the microsecond regime *routinely* we perform a microsecond long trajectory performed on a workstation-class PC. We use for this task the chicken Villin headpiece (HP-35) structure, one of the smallest polypeptides with a stable globular structure comprising three alpha helices placed in a "U"-shaped form, as shown in Figure 4c. Due to its small size, it is commonly used as a subject in long molecular simulations for folding studies, for instance,ref 43, which uses highly parallel distributed computing to compute many trajectories to fully sample the phase space of the folding process. Additionally, mutanogesis studies on fold-

Table 5. Performance of ACEMD on the DHFR, apoA1 Benchmark and Villin Test on 1 and 3 GPUs up to 720 Cores[a]

system	CPUs and GPUs	ns/day
DHFR	1 CPU, 1 GPU (240 cores)	19.7
DHFR	1 CPU, 3 GPUs (720 cores)	45.7
apoA1	1 CPU, 1 GPU (240 cores)	4.6
apoA1	3 CPU, 3 GPUs (720 cores)	10.6
Villin	3 CPU, 3 GPUs (720 cores)	66.0

[a] ACEMD run using Nvidia GTX 280 GPUs on real production runs (R = 9 Å, PME every step, time step 4 fs, constraints, and Langevin thermostat).

ing[44] have been performed using biased methods to accelerate the sampling.

The Villin headpiece (PDB:1YRF) was fully solvated in TIP3P water and Na−Cl at 150 mM (a total of 13 701 atoms) using the program VMD[45] and the CHARMM force field. The system was then equilibrated at 300 K and 1 atm for 10 ns using NAMD2.6[7] with a cutoff of 9 Å, PME with a 48 × 48 × 48 grid, constraints for all H bond terms, and a time step of 2 fs. Simulations with ACEMD were performed using an NVT ensemble, hydrogen mass repartitioning, and a time step of 4 fs. Starting from the final equilibrium configuration of NAMD, we run ACEMD at 450 K for 40 ns until the system was completely unfolded (movie available in ref 46). The resulting extended configuration (Figure 4b) was then used as the starting point of a microsecond long single trajectory at a temperature of 305 K. Figure 4a shows the rmsd of the backbone of the protein along the trajectory. The minimum rmsd was 4.87. The protein seems to sample quite often the overall shape of the crystal structure yet not converged toward it (Figure 4c). As this structure is expected to fold in 4−5 μs, we plan to extend the dynamics in the future along with any newer and faster version of ACEMD (for instance, using the new Nvidia GTX295 cards or, more likely, quad GPU Tesla S1075 units). An important consideration with regard to the force field is the following: with molecular simulations approaching microseconds, it is clear that the accuracy of the force fields will become more and more important. In particular, this system has been shown to be very sensitive to the force field used[44] (CHARMM seems to converge poorly toward the folded structure).

The production run on a PC equipped with ACEMD and 3 Nvidia GPUs (720 cores) required approximately 15 days (66 ns/day) (see Table 5) and probably represents the limit for current hardware and software implementation, while 5 μs should be obtainable in the near future using a 4-way GTX295-based system with 8 GPU cores. Using currently available commodity technology, the construction of computer systems with up to 8 directly attached GPUs has been demonstrated.[16,47] GPUs attach to the host system using the industry-standard PCI-Express interface.[48] This interface is characterized by a bandwidth comparable to that of main system memory (up to 8 GB/s for 16 lane PCIe 2.0 links typically used by graphics cards) but with a relatively higher latency.

The GPU resource requirements of the nonbonded kernel make it possible for up to 8 independent blocks to be processed simultaneously per multiprocessor. The limit of parallelization for the execution of the nonbonded kernel occurs when all blocks may be processed simultaneously by the available multiprocessors. Thus, for instance, a cubic simulation box with l = 66 Å and cell size 6 Å would scale over 167 multiprocessors (1336 cores) or 6 G200-class GPUs. Figure 5 shows the runtime of the inner and outer nonbonded kernels on a water box as a function of block count per kernel invocation. The minimum computation time for the fully parallel case would be 3.4 ms/step on current hardware. To further improve performance, optimization of the kernel or further subdivision of the computation would be required.

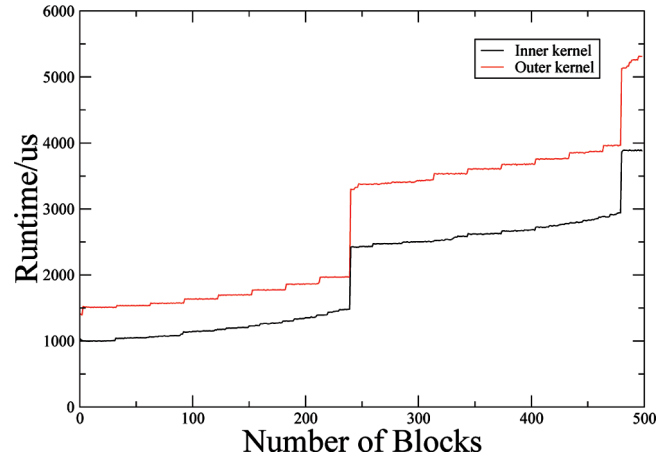


Figure 5. Run time for the nonbonded force calculation kernels on a water box as a function of the number of blocks per invocation and run on an Nvidia Tesla C1060 GPU (30 multiprocessors). The inner and outer kernels both have an occupancy of 8 blocks/multiprocessor. Blocks are distributed across multiprocessors, with the small step increases indicating an increment in the number of simultaneous blocks. The large steps indicate the device is fully populated with blocks and that some MPs must sequentially process further block. Optimal resource usage occurs immediately before these steps. The effect of gradual divergence between multiprocessors is seen as block count increases. The minimum run time for the fully parallel case would be 3.4 ms.

VII. Conclusions

We presented a molecular dynamics application, ACEMD, designed to reach the microsecond time scale even on cost-effective workstation hardware using the computational power of GPUs. It supports the CHARMM27 and Amber force field and is therefore suitable for use in modeling biomolecular systems. The ability to model these systems for tens of nanoseconds per day makes it feasible to perform simulations of up to the microsecond scale over the course of a few weeks on a suitable GPU-equipped machine. Calculations lasting a few weeks are perfectly reasonable tasks on workstation-class computers equipped with single or multiple GPUs.

ACEMD has been extensively tested since August 2008 through its deployment on the several thousand GPU-equipped PCs which participate in the volunteer distributed computing project GPUGRID.net,[49] based on the Berkeley Open Infrastructure for Networked Computing (BOINC)[50]

middleware. At the time of writing, GPUGRID.net delivers over 30 Tflops of sustained performance[51] and is thus one of the largest distributed infrastructures for molecular simulations, producing thousands of nanosecond long trajectories per day for high-throughput molecular simulations, for instance, for accurate virtual screening.[52]

The current implementation of ACEMD limits its parallel performance to just 3 GPUs due to a simple task parallelization. We plan to extend the use of ACEMD on more GPUs but keeping the focus on scalability, so small numbers of GPUs (1−32). Ideally the optimal system for ACEMD would rely on a single node attached to a large number of GPUs via individual PCIe expansion slots in order to take advantage of the large interconnect bandwidth. ACEMD would potentially scale very well on such a machine due to the fact that it is entirely executing on the GPU devices, obtaining CPU loads within just 5%. For efficient scaling across a GPU-equipped cluster, we anticipate that a refactoring of the parallelization scheme to use a spatial decomposition method[4] would be necessary, moving away from the simple task parallelization used in this work. Possible future developments also include support of forthcoming programming languages for GPUs, for example, OpenCL,[53] a development library which is intended to provide a hardware-agnostic, data-parallel programming model. While GPU devices are commonly present in desktop and workstation computers for graphics purposes, as accelerator processors they have yet to become routinely integrated components of the compute cluster systems typically used for high-performance computing (HPC) systems. GPU workstations, such as the one used in this work, are readily available, while GPU clusters are slowly appearing.[57] In order to scale efficiently, low-latency, high-bandwidth communications between nodes is necessary. For example, Bowers et al.[5] describe the scaling of the Desmond MD program over an Infiniband[54] network and demonstrate improved scaling when using custom communications routines tailored to the requirements of the algorithm and the capabilities of the network technology.

Accelerated molecular dynamics on GPUs as provided by ACEMD should be of wide interest to a large number of computational scientists as it provides performance comparable to that achievable on standard CPU supercomputers in a laboratory environment. Even research groups that have routine access supercomputing time might find useful the ability to run simulations locally for longer time windows and with added flexibility.

Acknowledgment. This work was partially funded by the HPC-EUROPA project (R113-CT-2003-506079). G.G. acknowledges support from the Marie Curie Intra-European Fellowship (IEF). G.D.F. acknowledges support from the Ramon y Cajal scheme and also from the EU Virtual Physiological Human Network of Excellence. We gratefully acknowledge the advice of Sumit Gupta (Nvidia), Acellera Ltd. (http://www.acellera.com) for use of their resources, and Nvidia Corp. (http://www.nvidia.com) for their hardware donations. We thank Ignasi Buch for help in debugging the software.

References

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

(2) Shaw, D. E. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Proceedings of the 34th annual international symposium on Computer architecture*; 2007.

(3) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2006**, *24*, 184109.

(4) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Phys.: Conf. Ser.* **2005**, *16*, 300–304.

(5) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *Proceedings of ACM/IEEE Conference on SuperComputing 2006*, Tampa, FL, Nov 11−17, 2006; ACM: New York, 2006.

(6) Fitch, B.; Rayshubskiy, A.; Eleftheriou, M.; Ward, T.; Giampapa, M.; Pitman, M.; Germain, R. IBM RC23956, May 2006, *12*.

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

(8) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theor. Comput.* **2008**, *4*, 435–447.

(9) De Fabritiis, G. *Comput. Phys. Commun.* **2007**, *176*, 600.

(10) Luttmann, E.; Ensign, D.; Vaidyanathan, V.; Houston, M.; Rimon, N.; Oland, J.; Jayachandran, G.; Friedrichs, M.; Pande, V. *J. Comput. Chem.* **2008**, *30*, 268.

(11) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903–1904.

(12) Harvey, M. J.; Giupponi, G.; Villà-Freixa, J.; De Fabritiis, G. PS3GRID.NET: Building a distributed supercomputer using the PlayStation 3. *Distributed & Grid Computing-Science Made Transparent for Everyone. Principles, Applications and Supporting Communities*; in press.

(13) Anderson, D. *Bekeley Open Infrastructure for Network Computing*; http://www.boinc.berkely.edu (accessed Apr 15, 2009).

(14) *OpenGL-The Industry Standard for High Performance Graphics*; Khronos Group: http://www.khronos.org/opengl/ (accessed Apr 15, 2009).

(15) Nickolls, J.; Buck, I.; Garland, M.; Skadron, K. *ACM Queue* **2008**, *6*.

(16) Stone, J.; Phillips, J.; Freddolino, P.; Hardy, D.; Trabuco, L.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.

(17) van Meel, J. A.; Arnold, A.; Frenkel, D.; Portegies-Zwart, S. F.; Belleman, R. G. *Mol. Simul.* 2008.

(18) Phillips, J. C.; Stone, J. E.; Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters. *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*; IEEE Press: 2008.

(19) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; LeGrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. **2009**, *30*, 864–872.

(20) Ufimtsev, I.; Martinez, T. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(21) Anderson, A. G.; Goddard, W. A., III; Schröder, P. *Comput. Phys. Commun.* **2008**, *177*, 298–306.

(22) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.

(23) Yasuda, K. *J. Chem. Theor. Comput.* **2008**, *4*, 1230–1236.

(24) Ewald, P. *Ann. Phys.* **1921**, *64*, 253–287.

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

(26) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(27) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: a portable plugin for free-energy calculations with molecular dynamics; 2009.

(28) Dongarra, J. J. Performance of various computers using standard linear equations software; Technical Report, Netlib report CS-89-85; 2007.

(29) Intel 64 and IA-32 Architectures Optimization Reference Manual, Document 248966-018, Technical Report; Intel: 2009.

(30) NVIDIA CUDA Compute Unified Device Architecture Programming Guide; Technical Report 2.0; NVIDIA Corp.: 2008.

(31) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *19*, 8577–8593.

(32) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.

(33) Hardy D. J. *MDX libraries*; University of Illinois at Urbana-Champaign: 2007; http://www.ks.uiuc.edu/Development/MDTools/namdlite.

(34) ACEMD website; http://multiscalelab.org/acemd (accessed Apr 15, 2009).

(35) CUDA CUFFT Library; Document PG-00000-003 V2.0, Technical Report; NVIDIA Corp.: 2008.

(36) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501–508.

(37) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.

(38) Verlet, L. *Part. Part. Syst. Charact.* **1967**, *159*, 98.

(39) Feenstra, K.; Hess, B.; Berendsen, H. *J. Comput. Chem.* **1999**, *20*, 786–798.

(40) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I. *J. Chem. Phys.* **2007**, *126*, 046101.

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

(42) Plimpton, S.; Hendrickson, B. *J. Comput. Chem.* **1996**, *17*, 326–337.

(43) Ensign, D.; Kasson, P.; Pande, V. *J. Mol. Biol.* **2007**, *374*, 806–816.

(44) Piana, S.; Laio, A.; Marinelli, F.; Van Troys, M.; Bourry, D.; Ampe, C.; Martins, J. *J. Mol. Biol.* **2008**, *375*, 460–470.

(45) Humprey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.

(46) Buch, I.; De Fabritiis, G. Movie of Vilin unfolding: http://www.vimeo.com/2505856 (accessed Apr 15, 2009).

(47) Vision Lab, U. FASTRA GPU SuperPC: http://fastra.ua.ac.be/en/index.html (accessed Apr 15, 2009).

(48) PCI Express Base Specification; Technical Report Revision 2.0; PCI Special Interest Group, 2007.

(49) De Fabritiis, G. The GPUGRID Project homepage: http://www.gpugrid.net (accessed Apr 15, 2009).

(50) Anderson, D. L. BOINC: A System for Public-Resource Computing and Storage. *Proceedings of Fifth IEEE/ACM International Workshop on Grid Computing (GRID'04)*; 2004.

(51) The BOINCStats Project statistics homepage: http://www.boincstats.com (accessed Apr 15, 2009).

(52) Holden, C., Ed. *Science* **2008**, *321*, 1425.

(53) OpenCL-The open standard for parallel programming of heterogeneous systems; Khronos Group: http://www.khronos.org/opengl/ (accessed Apr 15, 2009).

(54) InfiniBand Architecture Specification; Technical Report Release 1.2.1; Infiniband Trade Association: 2007. Available online at http://www.infinibandta.org/specs/register/publicspec/ (accessed Apr 15, 2009).

(55) The warp size is device specific and 32 for all current devices. As an implementation detail, each MP has 8 scalar cores and the warp threads are scheduled as 4 groups of 8 threads, with the execution of the groups pipelined on the scalar cores.

(56) This technique is also used in 18 and illustrates the general principal that, when programming the current generation of GPUs, the most efficent algorithms are those which favor simpler flow control and minimize global memory access even at the expense of redundant computation.

(57) Nvidia. Tesla-equipped Tsubame cluster; Tokyo Institute of Technology.

CT9000685

JCTC Journal of Chemical Theory and Computation

Quantum Cluster Equilibrium Theory Applied in Hydrogen Bond Number Studies of Water. 1. Assessment of the Quantum Cluster Equilibrium Model for Liquid Water

S. B. C. Lehmann, C. Spickermann, and B. Kirchner*

Wilhelm-Ostwald Institute of Physical and Theoretical Chemistry, University of Leipzig, Linnéstrasse 2, D-04103 Leipzig, Germany

Received July 31, 2008

Abstract: Different cluster sets containing only 2-fold coordinated water, 2- and 3-fold coordinated water, and 2-fold, 3-fold, and tetrahedrally coordinated water molecules were investigated by applying second-order Møller−Plesset perturbation theory and density functional theory based on generalized gradient approximation functionals in the framework of the quantum cluster equilibrium theory. We found an improvement of the calculated isobars at low temperatures if tetrahedrally coordinated water molecules were included in the set of 2-fold hydrogen-bonded clusters. This was also reflected in a reduced parameter for the intercluster interaction. If all parameters were kept constant and only the electronic structure methods were varied, large basis set dependencies in the liquid state for the density functional theory results were found. The behavior of the intercluster parameter was also examined for the case that cooperative effects were neglected. The values were 3 times as large as in the calculations including the total electronic structure. Furthermore, these effects are more severe in the tetrahedrally coordinated clusters. Different populations were considered, one weighted by the total number of clusters and one depending on the monomers.

1. Introduction

Calculating thermodynamic properties of condensed phases gives rise to substantial problems in computational chemistry, especially if systems exhibiting complicated electronic structures are involved; see e.g., refs 1 and 2. In general, these systems have to be treated in terms of quantum chemical *first-principles* methods, e.g., ab initio molecular dynamics (AIMD) simulations, from which reliable thermodynamic data for systems with many degrees of freedom are only obtained under very large computational efforts, if at all. The quantum cluster equilibrium (QCE) model circumvents these sampling problems of the phase space by decomposing the condensed phase into a thermodynamic equilibrium of distinct cluster structures, which in the zeroth approximation are treated as indistinguishable, noninteracting particles.[3,4] For this noninteracting cluster phase an analytical, ideal-gas-like partition function is available, which gives direct access to the thermodynamics of the system. Two central corrections are introduced to this ideal "cluster gas" to account for the special conditions at liquid-phase densities, namely, the reduced free volume of translation and the interaction between different clusters, which are adjustable by means of two scaling parameters. The cluster structures and corresponding properties are obtained from static quantum chemical calculations, in which sophisticated ab initio methods are applicable, thus enabling the treatment of demanding electronic structures on the cluster level. The intercluster interaction is realized in the present model according to a nonlocal, van der Waals-like mean field potential, which in principle could be replaced by more advanced expressions. In this way the QCE approach introduces ab initio quantum chemistry including correlated electronic structure methods to the condensed phase. A recent example for this procedure is the determination of the vaporization entropy of water, which for the first time has

* Corresponding author phone: 493419736401; fax: 493419736399; e-mail: bkirchner@uni-leipzig.de.

10.1021/ct800310a CCC: $40.75
Published on Web 05/13/2009

been calculated on the basis of correlated electronic structure methods.[5] Besides the accurate treatment of electron correlation, nuclear quantum effects are within the scope of the model in terms of ab initio cluster frequency analyses as well. However, the maybe most important point concerning the investigation of highly associated liquids is the inclusion of cooperative effects, which have been demonstrated to be of high importance for the accurate calculation of various thermodynamic properties for the liquid water phase.[5,6]

The QCE model already proved to be applicable to a variety of associated liquids, the most prominent of which is pure water. The first QCE application already pointed out that in the case of water generic structural patterns (so-called "cluster motifs") are more important than special cluster sizes and geometries, and additional studies demonstrated that the model is capable of reproducing large parts of water's phase diagram at least qualitatively, including the triple point and a phase transition to an ice-equivalent solid phase.[7,8] Additional examinations of water in terms of the QCE approach include investigations on isotopically substituted water as well as the influence of quantum chemical methodology on the density of the liquid state.[4,6,9] Besides pure water, the QCE model was furthermore successfully applied to different other associated liquids, for instance, formamide, methanol, ethanol, formic acid, and liquid sulfur, to name but a few.[10−15] More recent investigations concern the importance of the tetrahedral coordination in liquid water for the reproduction of its anomalous properties as well as the influence of cooperative and dispersion effects in liquid *cis*,*cis*-cyclotriazane.[16,17] These studies suggest that the cluster approach is a reliable approximation to the thermodynamics of the condensed phase, at least in the case of highly associated liquids. As already mentioned, the two most crucial points in the QCE procedure are the approximate treatments of intercluster interaction and excluded volume. Due to these corrections, the aforementioned ideal cluster gas in a sense becomes a real (that is condensed) cluster gas, in which the constituents exhibit an appropriate volume and are allowed to interact with each other. The isolated cluster structures obtained from the static quantum chemical calculations are embedded in an attractive mean field potential, which models the dense character of associated liquids and which is the reason why QCE calculations clearly have to be distinguished from cluster studies in which isolated cluster entities are applied. Thus, the present QCE approach can be understood as the first step toward the answer to the legitimate question of what clusters can tell us about the condensed phase.

In the present study we introduce only slightly larger water clusters compared to those investigated previously[4,6] to pronounce the effect due to the tetrahedral coordination and not due to more compact or much larger structures. However, these new clusters contain rings interconnected via a tetrahedrally coordinated water molecule, and we present results of electronic structure calculations with explicit correlation, i.e., Møller−Plesset perturbation theory (MP2). The Results start with improved accuracies for the old (**2/3**) cluster set formerly denoted as the **7(w8cube)** cluster set.[4,6] This part is followed by a comparison of isobars including and excluding the new tetrahedrally coordinated water molecules. We determine optimal cluster sets with respect to the calculated isobars ($\mathbf{2_{opt}}$ = old optimized and $\mathbf{2-4_{opt}}$ = new optimized) by deleting underpopulated clusters. The optimization of a cluster set as the basis for the QCE calculation is thereby examined in detail. Furthermore, we analyze the liquid-phase composition in terms of monomer-normalized populations. Next, we study dispersion versus cooperative effects at constant parameters, and we discuss the different water model structures and the shortcomings of theoretical investigations. The present paper ends with the Conclusion.

2. Methodology

2.1. Quantum Cluster Equilibrium Details. A full derivation of the QCE theory can be found elsewhere.[3,4,6] The most important aspects of the QCE method are given in the following.

2.1.1. Partition Functions. Neglecting vibrational−rotational interactions and other small perturbations, the cluster partition function can be factorized into the translational ($q_{j,\mathrm{trans}}$), vibrational ($q_{j,\mathrm{vib}}$), rotational ($q_{j,\mathrm{rot}}$), and electronic ($q_{j,\mathrm{elec}}$) contributions in the usual way, resulting in

$$q_j = q_{j,\mathrm{trans}} q_{j,\mathrm{vib}} q_{j,\mathrm{rot}} q_{j,\mathrm{elec}} \tag{1}$$

In the high-temperature continuum limit the translational partition function is given by

$$q_{j,\mathrm{trans}} = \frac{V - V_{\mathrm{excl}}}{\Lambda_j^{\,3}} \tag{2}$$

Here Λ_j is the thermal de Broglie wavelength in one dimension

$$\Lambda_j = \frac{h}{(2\pi m^{(j)} k_{\mathrm{B}} T)^{1/2}} \tag{3}$$

with $m^{(j)}$ being the mass of cluster j, h the Planck constant, T the temperature, and k_{B} the Boltzmann constant. The numerator of eq 2 describes the available volume for free translational motion. This free volume is obtained by subtracting an excluded volume proportional to the total molecular volume of all clusters from the phase volume

$$V_{\mathrm{exc}} = b_{\mathrm{xv}} \sum_{j=1}^{\eta} n_j V_j \tag{4}$$

where the proportionality constant b_{xv} serves as one of two variable parameters to adjust the calculations to experimental data. The rotational partition function $q_{j,\mathrm{rot}}$ is also deduced from the continuum limit as

$$q_{j,\mathrm{rot}} = \frac{1}{\sigma}\left(\frac{\pi T^3}{\Theta_{\mathrm{A}}\Theta_{\mathrm{B}}\Theta_{\mathrm{C}}}\right)^{1/2} \tag{5}$$

with the rotational symmetry factor σ derived from the optimized cluster structure and the rotational temperatures represented by Θ_{A}, Θ_{B}, and Θ_{C}. These can be calculated from the three principle rotation axes achieved from the

calculated moments of inertia I_A, I_B, and I_C of a given cluster by following the equation

$$\Theta_X = \frac{\hbar^2}{2I_X k_B} \tag{6}$$

with X = A, B, or C. The next term gives the vibrational contribution $q_{j,vib}$. For each of the $3N - 6$ normal modes, with N being the number of atoms in the molecule, the harmonic oscillator approximation is employed, resulting in

$$q_{j,vib} = \prod_{n=1}^{3M_j-6} (e^{-\Theta_n^{(j)}/2T})(1 - e^{-\Theta_n^{(j)}/T})^{-1} \tag{7}$$

with $M_j = 3i_j a$ as the number of atomic nuclei in cluster j with a atoms per monomer i and the vibrational temperature represented by $\Theta_n^{(j)}$

$$\theta_n^{(j)} = \frac{h\nu_n^{(j)}}{k_B} \tag{8}$$

which is associated with the vibrational frequency $\nu_n^{(j)}$ of the nth normal mode. In eq 7 the zero-point vibrational energy (ZPVE) is taken into account by the $e^{-\theta_n^{(j)}/2T}$ term. The treatment of this nuclear quantum effect is not included in every method; for example, most molecular dynamics simulations neglect this contribution.[18]

The last contribution to the canonical partition function is a modified electronic part, depending on the cluster interaction energies of the different species.[3] Therefore, the zero point of the energy scale is set to the total ground-state energy of the relaxed monomer E_1. The cluster interaction energy includes on one hand the important nonpairwise additive cooperative effects and on the other hand, according to the used electronic structure method, dispersion effects. Additionally, pairwise additive interaction energies obtained in the way described in ref 6 were applied to this contribution instead of the total cooperative energies. At this point it should be mentioned that the cluster interaction energies ($\Delta E_j^{cp} = E_j^{cp} - i_j E_1$) were always corrected by employing a full counterpoise correction as introduced by Boys and Bernardi.[19]

Here higher contributions than the electronic ground state are neglected. Up to this point only intracluster interaction energies were taken into account. To treat the attractive intercluster interaction energies as well, the volume- and cluster-size-dependent mean field potential energy

$$E_j^{inter} = -i_j a_{mf} V^{-1} \tag{9}$$

was introduced to the model, containing the mean field parameter a_{mf}.[3] Finally, the complete electronic contribution to the partition function reads

$$q_{j,elec} = e^{[\Delta E_j^{cp} + E_j^{inter}]/k_B T} \tag{10}$$

2.1.2. Equilibrium and the Polynomials. The QCE method assumes a thermodynamic equilibrium between differently sized clusters and one corresponding reference cluster. The largest possible number of monomers ξ_η of the largest cluster X_n in this equilibrium may be different from the total number of clusters η. Following from this, the number of monomers i_j in a given cluster and the "list" number j of that cluster can be different. Thus, the equilibrium reads

$$X_1 \leftrightarrows \frac{X_2}{2} \leftrightarrows \frac{X_3}{3} \leftrightarrows \frac{X_4}{4} \leftrightarrows \ldots \leftrightarrows \frac{X_j}{i_j} \leftrightarrows \ldots \leftrightarrows \frac{X_\eta}{\xi_\eta} \tag{11}$$

Here X_j denotes a cluster of i_j monomer units up to ξ_η monomer units forming the largest cluster. η represents the total number of clusters. Using the relation between the cluster partition function q_j and the chemical potential μ_j for which the same equilibrium holds as in eq 11

$$\mu_j = -k_B T \ln\left(\frac{q_j}{N_j}\right) \tag{12}$$

and assuming particle conservation

$$N_A = N_1 + 2N_2 + 3N_3 + 4N_4 + \ldots + i_j N_j + \ldots + \xi_\eta N_\eta \tag{13}$$

leads per insertion of eq 12 into eq 13 to an iterative cycle for the root finding of the population polynomial and the volume polynomial.[6] N_A denotes the Avogadro number and N_j the particle number. Changing to moles $n_j = N_j/N_A$ instead of particle numbers, the population polynomial for the monomer (with the partition function q_1) is given by the following expression:

$$0 = -1 + \sum_{j=1}^{\eta} \left[\frac{i_j q_j (N_A)^{i_j - 1}}{(q_1)^{i_j}}\right] (n_1)^{i_j} \tag{14}$$

In a similar manner one arrives at the volume polynomial:

$$0 = -pV^3 + [RT\sum_{j=1}^{\eta} n_j + pV_{excl}]V^2 - [\sum_{j=1}^{\eta} i_j n_j a_{mf}]V + [\sum_{j=1}^{\eta} i_j n_j a_{mf}]V_{excl} \tag{15}$$

R is the ideal gas constant, and p is the chosen pressure. The degree of the volume polynomial depends on whether the mean field interaction is employed ($a_{mf} \neq 0$) or not ($a_{mf} = 0$). From quantum chemical calculations the input for the partition functions is obtained. These will be described in the next section in brief. The partition functions will next serve in the polynomial equations, which leads to different sets of populations and volumes. From these sets the one combination with minimal Gibbs energy is chosen, and this volume/population combination is re-entered into the iterative cycle until volume convergency is reached.

2.2. Computational Details. It is important to repeat that inherent to the QCE model we define two interaction terms: First is the *intercluster* interaction, which is the interplay between different clusters; see also section 2.1.1. In the QCE model this interaction is accounted for by a mean field energy term depending on the mean field parameter a_{mf}.[3,7] Second is the *intracluster* interaction, which represents the binding energy of a single cluster and which is

$$\Delta E_j^{cp} = E_j^{cp} - i_j E_1 \quad (16)$$

where E_j and E_1 denote the total energies of the jth cluster containing i_j monomers and the corresponding monomer in its relaxed geometry (adiabatic interaction energy). It is obvious that the *intercluster* term containing a_{mf} accounts for the deficiencies of the cluster approach due to incomplete solvation and that both terms contribute to the condensed-phase behavior.

The pair energies were calculated as described previously.[4,6] All pair energies listed in ref 6 were recalculated. Here only the interaction between each pair in a cluster is considered. Structure optimizations were performed employing density functional theory (DFT) as well as second-order Møller−Plesset perturbation theory (MP2) with the resolution of identity (RI) procedure.[20] The program packages used for the electronic structure calculations were Turbomole 5.91 and associated programs.[20] For DFT calculations, the gradient-corrected functional BP86 was employed in combination with the TZVP and TZVPP basis sets as well as the RI technique. The MP2 calculations were additionally carried out with the TZVP and TZVPP basis sets.[20] As stated above, the basis set superposition error is treated in terms of the counterpoise correction of Boys and Bernardi.[19] For the determination of the principal moments of inertia and the harmonic frequencies, the SNF program package was employed after the electronic structure calculations were carried out.[21] The SNF program computes frequencies on the basis of the harmonic approximation as numerical derivatives of the analytic gradients provided by the structure optimization routine. All harmonic frequencies enter the vibrational partition function unscaled. The two parameters of the QCE model are adjusted to reproduce experimental volumes[22] only and not to reproduce other quantities. Once the parameters are chosen, they are fixed for the calculation of other quantities.

The QCE calculations were performed employing the PEACEMAKER code.[23] To obtain optimal values for the excluded volume and mean field interaction parameters, a sampling of isobars over a predefined a_{mf}/b_{xv} interval is carried out, and the "best choice" isobar with respect to the experimental curve is determined. The employed selection procedure is a straightforward application of the commonly used least-squares fit and is described elsewhere.[5] Within this work we adjusted a_{mf} and b_{xv} to achieve the accuracy $||\Delta V||$ to the fourth decimal place, which equates to a magnitude of microliters.

3. Cluster Sets Investigated

To conduct our study on liquid water, we used the cluster set as introduced in the first publication of Weinhold[3,7] and as employed previously by the present authors;[4−6] see Figure 1. The original **2/3** cluster set (formerly denoted as the **7(w8cube)** set)[4−6] mainly contains structural motifs of a 2-fold coordinated water molecule with as many acceptor−donor (AD) hydrogen bonds as water molecules in the cluster. Only in the **w8cube** (which replaced the ringlike **w8** cluster,[4−6] because the **w8** ring was not found to be a minimum for the MP2 method) is at least a 3-fold coordination provided with four ADD and four AAD coordinated molecules.

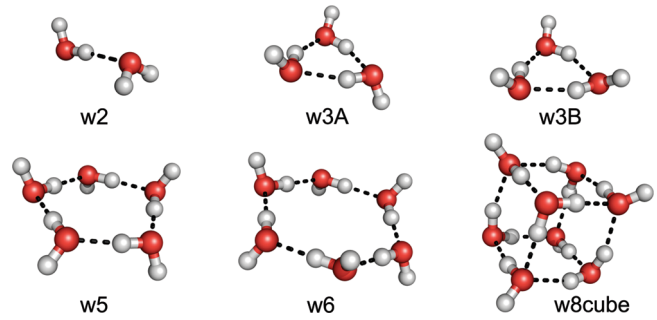


Figure 1. Ball-and-stick model of the original cluster set (abbreviated as the **2/3** set) as introduced by Weinhold[3,7] and as employed previously by the present authors.[4−6]

Table 1. Adiabatic Interaction Energies ΔE_j and Negative Energies per Monomer $-\Delta E_j/i_j$ (kJ/mol) from Refs 4 and 6[a]

	BP		MP2		
	TZVP	TZVPP	TZVP	TZVPP	
cluster j	cooperative	cooperative	cooperative	cooperative	pair
			ΔE_j		
w2	−20.9	−18.0	−19.9	−19.2	−19.0
w3A	−68.5	−61.4	−60.3	−61.7	−46.4
w3B	−63.9	−57.3	−57.3	−58.6	−44.0
w5	−162.3	−147.9	−140.4	−140.9	−68.2
w6	−201.4	−182.5	−175.2	−175.0	−74.3
w8cube	−311.9	−283.0	−269.9	−280.9	−98.4
			$-\Delta E_j/i_j$		
w2	10.5	9.0	10.0	9.6	9.5
w3A	22.8	20.5	20.1	20.6	15.5
w3B	21.3	19.1	19.1	19.5	14.7
w5	32.5	29.6	28.0	28.2	13.6
w6	33.6	30.4	29.2	29.2	12.4
w8cube	39.0	35.4	33.7	35.1	12.3
			$-\Delta E_j/n^{hb}$		
w2	20.9	18.0	19.9	19.2	19.0
w8cube	26.0	23.6	22.5	23.4	8.2

[a] The last two lines give the negative energy per hydrogen bond (n^{hb}), which is different for the dimer and the **w8cube** cluster only.

In Table 1 we list the interaction energies and the negative energies per monomer for the **2/3** cluster set. For all methods and basis sets employed **w8cube** is the most stable cluster followed by the **w6** and **w5** clusters. In the pair energies per monomer (second block, last column, Table 1) these trends are not present. Although **w2** is still the least stable cluster, it is followed by the **w8cube** and **w6** clusters. Thus, in **w6** and **w8cube** we observe large cooperative effects.

To probe the 4-fold coordination, further clusters were added; see Figure 2. We call these additional clusters "spiro clusters'' to illustrate the analogy to organic spiro compounds. A further feature of the spiro clusters next to the structural motif of the 4-fold coordination is that they are larger than most of the clusters from the **2/3** set. The leading structural motif in all spiro clusters is one AADD water molecule and $i_j - 1$ molecules with AD hydrogen bonds per water molecule. The combination of these new spiro-type clusters and the members of the old **2/3** set will be called **2−4** set. The interaction energies together with the energies per monomer and per hydrogen bond and the basis set

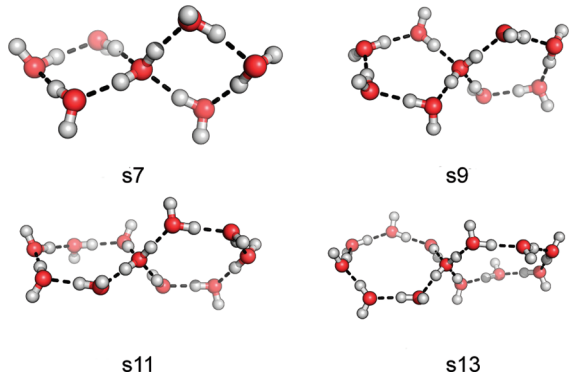


Figure 2. Ball-and-stick model of additional clusters.

Table 2. Adiabatic Interaction Energies ΔE_j, BSSEs, Negative Energies per Monomer $-\Delta E_j/i_j$, and Negative Energies per Hydrogen Bond $-\Delta E_j/n^{hb}$ (kJ/mol)

	BP		MP2		
	TZVP	TZVPP	TZVP	TZVPP	
cluster j	cooperative	cooperative	cooperative	cooperative	pair
			ΔE_j		
s7	−247.3	−218.6	−203.1	−208.8	−90.6
s9	−324.7	−296.5	−279.9	−281.7	−75.0
s11	−392.5	−356.3	−342.5	−342.8	−66.5
s13	−461.6	−416.6	−405.2	−402.8	−52.1
			BSSE		
s7	−24.6	−24.6	−56.0	−40.4	−307.6
s9	−44.1	−31.5	−75.4	−53.4	−101.3
s11	−54.3	−38.4	−91.4	−63.8	−102.8
s13	−62.2	−43.7	−104.8	−72.6	−212.8
			$-\Delta E_j/i_j$		
s7	35.3	31.2	29.0	29.8	12.9
s9	36.1	32.9	31.1	31.3	8.3
s11	35.7	32.4	31.1	31.2	6.0
s13	35.5	32.1	31.2	31.0	4.0
			$-\Delta E_j/n^{hb}$		
s7	30.9	27.3	25.4	26.1	11.3
s9	32.5	29.7	28.0	28.2	7.5
s11	32.7	29.7	28.5	28.6	5.5
s13	33.0	29.8	28.9	28.8	3.7

superposition errors for different methods and basis sets are listed in Table 2.

Employing a larger basis set (TZVPP instead of TZVP) leads to a decrease of the absolute values of the interaction energies for the density functional methods and to an increase of the absolute values of interaction energies in the case of MP2, with the exception of **s13**. The basis set superposition errors (BSSEs) are in the range of 20−100 kJ/mol, depending on the method and basis set as well as the cluster size. As can be expected, the BSSEs decrease with increasing basis set. Considering the energies per monomer, we find that within a chosen basis set and method the energies are all similar, i.e., around 36 kJ/mol for BP/TZVP, 32 kJ/mol for BP/TZVPP, and around 31 kJ/mol for the MP2 calculations. Thus, these clusters are all more stable than those from the **2/3** cluster set with the exception of the **w8cube**; see Table 1. Again the pair energies reverse trends, showing that large cooperative effects do play a role. Due to their geometrical conformation, the spiro clusters show a higher cooperativity on average; i.e., they have smaller pair energies than the clusters from the old **2/3** set. Considering the energy per hydrogen bond, it is obvious why the **w5** and **w6** clusters are such important structures for the liquid phase, because these clusters are those with the strongest hydrogen bonds on average. However, the spiro clusters are only a little less stable than the **w6** cluster. As can be observed from the last block of Table 2, these kinds of clusters provide average hydrogen bonds as strong as those of the **w5** cluster and almost as strong as those of the **w6** cluster. In general, we observe the known overbinding of DFT;[24] e.g., for small basis sets BP always leads to stronger bound clusters.

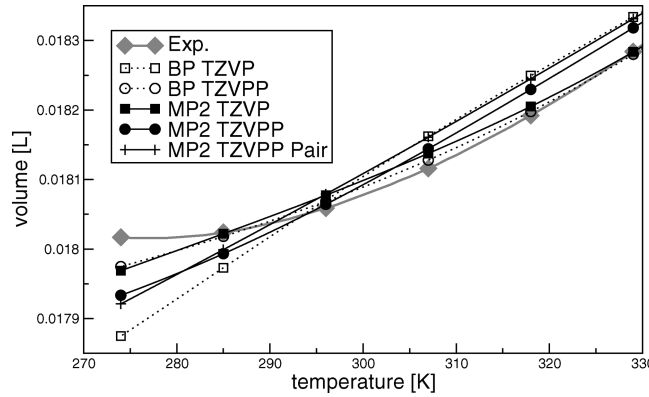


Figure 3. Calculated isobars for the **2/3** cluster set displayed at a temperature range of 274−329 K. For each electronic structure method a_{mf} and b_{xv} are newly adjusted (range 274−373 K); for the results see Table 3.

The corresponding frequency calculations contain only positive values, indicating that we are dealing with minimum structures. For the sake of reproducibility these data can be obtained upon request from the authors.

4. Results

4.1. 2/3 Set: Isobars Revisited and Optimal Set. We start this section with the newly obtained isobars for the original **2/3** cluster set, i.e., without applying the spiro clusters. The isobars are illustrated in Figure 3. In this figure two electronic structure methods (BP, MP2) and two basis sets (TZVP, TZVPP) as well as MP2/TZVPP pair energies were considered for the selection procedure. It can be recognized that BP/TZVPP follows the experimental curve closest, while BP/TZVP leads to the worst agreement with the experimental data.[22] The MP2 data represent a more accurate electronic structure, because the method treats the electron−electron correlation in higher detail than density functional theory based on the generalized gradient approximation. This does not necessarily lead to the best isobars when the model-inherent parameters of the QCE calculations are allowed to vary. Thus, it is difficult to discuss trends at "improved" electronic structure methods. Interestingly, we obtain different trends for the varying electronic structure methods if we compare the low temperature range and high temperature range. Furthermore, it is evident that the deviations from the experimental isobar at lower temperatures are larger for all methods than at higher temperature. From this behavior the first implications of the importance of the tetrahedral coordination pattern at lower temperatures can be drawn.

Table 3. QCE Parameters for Different Methods and Basis Sets without Spiro Clusters (**2/3** Set) in the Temperature Range from 274 to 373 K[a]

method	basis set	$\|\|\Delta V\|\|$	a_{mf}	b_{xv}
BP	TZVP	523.56	0.121	1.084
BP	TZVPP	135.49	0.158	1.110
MP2	TZVP	175.05	0.162	1.099
MP2	TZVPP	311.80	0.162	1.107
MP2 Pair	TZVPP	445.76	0.370	1.078

[a] $||\Delta V||$ (μL) gives the accuracy of the selected isobar, i.e., the root mean square deviation from the experimental values.

Table 4. QCE Parameters for MP2/TZVPP Calculations without Spiro Clusters (**2/3** and $\mathbf{2_{opt}}$) and with the Spiro Clusters (**2**−**4** and $\mathbf{2-4_{opt}}$) in the Temperature Range from 274 to 373 K[a]

set	$\|\|\Delta V\|\|$	a_{mf}	b_{xv}
2/3	311.80	0.162	1.107
$\mathbf{2_{opt}}$	167.57	0.136	1.088
2−**4**	166.42	0.125	1.105
$\mathbf{2-4_{opt}}$	137.07	0.130	1.105

[a] $||\Delta V||$ (μL) gives the accuracy, i.e., the root mean square deviation, from the experimental values.

In the present work we selected the QCE parameters with one more decimal place as in previous studies, which improves in some cases the accuracy by 50%.[5,6] In Table 3 we show the accuracies of the selected QCE parameters for the different electronic structure methods used within this work. The same trends in the value of the accuracy $||\Delta V||$ as those discussed for the plots of Figure 3 can be observed in Table 3. A highly intuitive point of the QCE model is that the smaller the intracluster interaction energies are, the larger the parameter accounting for the mean field interaction a_{mf} becomes. This demonstrates that the parameter in a way accounts for the deficiencies of the missing interactions, whether arising from the intercluster part or not treated correctly by the electronic structure method (for instance, overestimated for BP/TZVP). This point is most striking for the pair energies where the a_{mf} value is more than twice as large as the a_{mf} value of the corresponding MP2/TZVPP QCE calculation. We will discuss this topic of neglected cooperativity in a later section. a_{mf} deviates from that of the uncorrected model ($a_{mf} = 0$ and $b_{excl} = 1$) by up to 0.37 which correspond to 37%, and b_{xv} only varies within 11%, again indicating the pronounced importance of the overall interparticle interactions for the treatment of condensed-phase phenomena.

4.1.1. Pure 2-Fold Coordinated Cluster Set and Monomer-Normalized Populations. To examine the influence of the different clusters in the set, we determined an optimal set by deleting clusters systematically under the proposition of keeping the same level of accuracy. Initially, we studied the elimination of those clusters that are found to be weakly populated, which led to the exclusion of the **w3B** cluster. Because the aim of the accompanying paper (10.1021/ct900189v) is to investigate the importance of the 2-fold versus the 4-fold coordination, we additionally constructed a set consisting of pure 2-fold hydrogen-bonded motifs. This means that **w8cube** is deleted from the cluster set as well, which leads to an even more accurate isobar for the MP2/TZVPP data, as can be seen in the first two lines of Table 4. The set obtained in this way will be denoted as the $\mathbf{2_{opt}}$ set. The resulting QCE parameters are $a_{mf} = 0.136$ and $b_{xv} = 1.088$, and the accuracy is $||\Delta V|| = 167.57$ μL, which is even smaller by a factor of 2 as compared to the accuracy of the complete **2/3** set; see Table 3. Interestingly, the a_{mf} value is reduced, which indicates that either the **w3B** or the **w8cube** cluster must have a destabilizing effect on the intercluster interaction. It also indicates that it is mandatory to optimize the cluster set as a basis for the QCE calculation to obtain excellent results. The old (gray) and newly obtained (black) cluster populations are shown on the left side of Figure 4. While only slight changes in the high-temperature region can be recognized (see the gray curves with squares in Figure 4, left panel) the **w6** cluster becomes the most important cluster at lower temperature (see the black curves with circles in Figure 4, left panel). The **w5** cluster population also increases a little bit at lower temperature as compared to the behavior in the **2/3** set. This is in accordance with the isolated molecule energetics formerly discussed in section 3. On the right-hand side of Figure 4 we show the monomer-normalized populations which will be used in the accompanying paper (10.1021/ct900189v), because they provide a more physical picture of the particular phase point. The cluster populations on the left side of Figure 4 refer to the total number of clusters composing the actual phase point, but this amount varies within each step of the QCE calculation. This means that if mainly large clusters are populated, more monomers are bound within these large clusters. Because the amount of monomers is fixed in the QCE calculation, there are fewer monomers to form the other clusters. This implies a reduction of the total number of clusters; i.e., if the total number of clusters decreases, the population of each individual cluster increases. Thus, for our purpose it is better to analyze monomer-normalized populations (see the right side of Figure 4), which are applied in the monomer reference system. For all phase points the total number of monomers is equal to 1 mol. The percentage now indicates how many of the total 1 mol monomers are bound in a particular cluster and thereby reflects the physical composition of the phase point. From Figure 4, right panel, we observe that the liquid phase of the $\mathbf{2_{opt}}$ set is almost completely composed of the **w6** cluster and the **w5** cluster. While the **w6** population decreases from 79% to approximately 63%, the **w5** population grows from 20% to 29% with increasing temperature.

4.2. 2−4 Cluster Set and Optimal $\mathbf{2-4_{opt}}$ Cluster Set. We now turn to the larger cluster set containing tetrahedrally coordinated water molecules, namely, the **2**−**4** cluster set. The obtained parameters and accuracies are listed in Table 4. The data clearly demonstrate that the inclusion of the spiro clusters leads to an improvement in the accuracy over that of the **2/3** cluster set. Interestingly, the a_{mf} value of the **2**−**4** set is reduced as compared to that of the **2/3** set. This is not a general rule; see Table 3. Here a better accuracy goes along with smaller and in some cases with larger a_{mf} values. However, from this observation we can deduce that the tetrahedrally coordinated water molecule in the particular

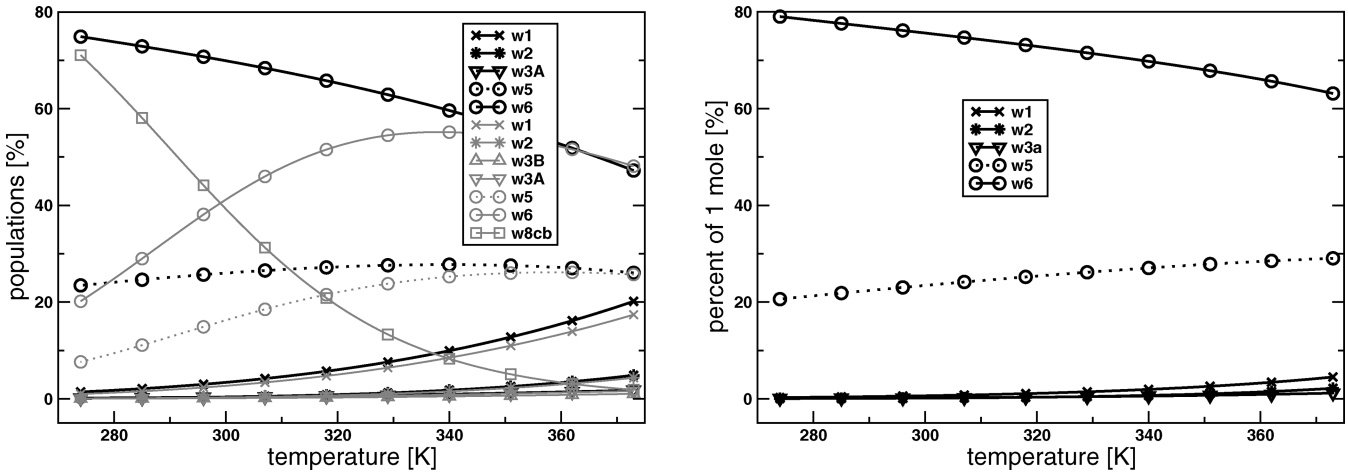


Figure 4. Obtained populations for the **2/3** cluster set (gray) and the $\mathbf{2}_{opt}$ cluster set (black) at the temperature range of 274−373 K for MP2/TZVPP: left, cluster populations; right: monomer-normalized populations. The QCE parameters are $a_{mf} = 0.136$ and $b_{xv} = 1.088$.

spiro cluster mimics the intercluster interaction more accurately, leading to a reduction of the a_{mf} value.

To examine the influence of the different clusters in the **2**−**4** set, we determine another optimal set (abbreviated as the $\mathbf{2}-\mathbf{4}_{opt}$ set) by systematically deleting clusters under the proposition of keeping the same level of accuracy. Therefore, we studied the elimination of those clusters that were found not to be populated highly, i.e., **w3A**, **w3B**, **s7**, and **s13**. Their elimination leads in some cases to similar results, but in nearly half the possibilities to slightly improved results as compared to taking the whole cluster set into account, which is shown in Table 6 in the Appendix. Thus, the $\mathbf{2}-\mathbf{4}_{opt}$ cluster set contains the old clusters **w1**, **w2**, **w3B**, **w5**, **w6**, and **w8cube** as well as the new **s9** and **s11** clusters. In Table 4 the results of the selection routine are also given for the $\mathbf{2}-\mathbf{4}_{opt}$ set. Again a_{mf} is reduced compared to that of the $\mathbf{2}_{opt}$ set, although only to a small extent. The QCE parameters are almost identical to those obtained for the complete set; see Table 4. What is also apparent from Table 4 is that the addition of clusters containing tetrahedrally coordinated water leads to a strong improvement of the accuracy; compare values for the **2/3** set with those for the **2**−**4** set or values for $\mathbf{2}_{opt}$ with those for the $\mathbf{2}-\mathbf{4}_{opt}$ set. Both quantities, the accuracy and a_{mf}, point to a better description of liquid water when tetrahedrally coordinated molecules are present. This will be discussed in more detail in the accompanying paper (10.1021/ct900189v).

4.3. Constant Parameters and Varying Electronic Structures. After definition of an optimal cluster set ($\mathbf{2}-\mathbf{4}_{opt}$), isobars are calculated at constant QCE parameters for all used electronic structure methods and their corresponding frequencies in this section. The results are shown in Figure 5. At higher temperature larger deviations from the experimental isobar can be observed. Again there are trends for the different electronic structure methods. The BP method with TZVP and TZVPP yields the highest intracluster interaction energies, which is reflected in small volumes. However, according to the larger energies found for BP/TZVP, smaller volumes than for the BP/TZVPP calculations would be expected, but the opposite trend is observed. This clearly shows the important influence of the other two

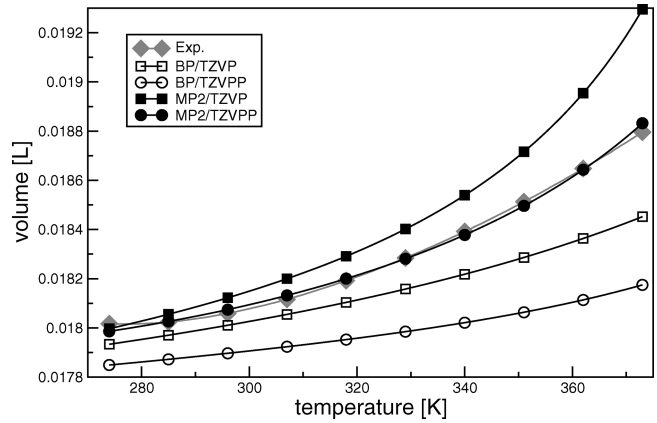


Figure 5. Isobars at the liquid-phase temperature range for the $\mathbf{2}-\mathbf{4}_{opt}$ set with $a_{mf} = 0.130$ and $b_{xv} = 1.105$.

partition functions, i.e, the rotational and the vibrational contributions, which in some cases disturb the reciprocally proportional relation between interaction energy and volume. According to Table 2, MP2/TZVP calculations lead to the smallest intracluster interaction energies; thus, the volumes are much larger than for the other methods.

To study the pure influence of the electronic structure method and the electronic partition function, we computed isobars at constant input for the rotational and vibrational partition functions and applied constant QCE parameters; see Figure 6. Because the MP2/TZVPP set parameters were adjusted to the experimental curve, we observe an excellent agreement between MP2/TZVPP and the experimental isobar. The MP2/TZVP curve deviates now much less as compared to the results depicted in Figure 5, indicating again the strong influence of the vibrational and rotational partition functions.

All methods show larger deviations at higher temperature. The smallest overall volumes are obtained from the BP/TZVP energies, which predict the strongest binding energies for all clusters compared to the other electronic structure methods. The BP/TZVPP isobar is slightly closer to the experiment. This reflects an interesting and unexpected basis set dependency of DFT and might be important in light of

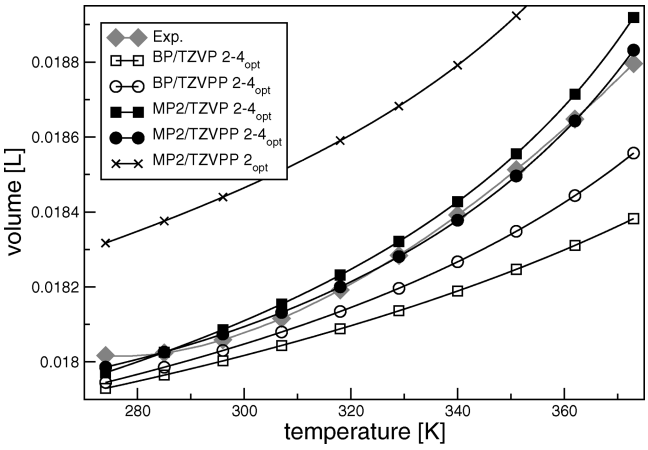


Figure 6. Isobars at the liquid-phase temperature range for the $\mathbf{2-4_{opt}}$ set with $a_{mf} = 0.130$ and $b_{xv} = 1.105$ at constant MP2/TZVPP frequencies and moments of inertia from MP2/TZVPP geometries.

Table 5. QCE Parameters for the TZVPP Basis Set and MP2 Calculations without Spiro Clusters ($\mathbf{2_{opt}}$) and with Spiro Clusters ($\mathbf{2-4_{opt}}$) in the Temperature Range from 274 to 373 K[a]

set	$\|\|\Delta V\|\|$	a_{mf}	b_{xv}
$\mathbf{2_{opt}}$	167.57	0.136	1.088
$\mathbf{2_{opt}}$ pair	405.63	0.343	1.068
$\mathbf{2-4_{opt}}$	137.07	0.130	1.105
$\mathbf{2-4_{opt}}$ pair	420.18	0.351	1.071

[a] $\|\|\Delta V\|\|$ (μL) gives the accuracy, i.e., the root mean square deviation from the experimental values.

the current basis set limit studies of Lee and Tuckerman, who found no glassy or overstructured state in a 60 ps first-principles simulation on water employing the BLYP density functional with a complete discrete variable representation (DVR) basis set.[24] The improved volume at larger basis set in the isobars of the density functional theory compared with the too small volume with small basis set might thus be in accordance with the "unfrozen water state" of the DVR simulation by Lee and Tuckerman.[24] Another interesting feature is the nonconstant deviation between calculated BP data and experimental data. At higher temperature the deviations become much larger due to the wrong slope of the BP isobars. Thus, our data indicate that using density functional theory (BP) as well as a small basis set leads to a less accurate description of the liquid phase especially at high temperature as compared to MP2 with the TZVPP basis set. Because we observe volumes which are too small for the isobars, we can deduce that both approximations could be the origin for an overstructured liquid as discussed by Lee and Tuckerman.[24]

Next, we compare the different monomer-normalized populations of the $\mathbf{2-4_{opt}}$ set, which are depicted in Figure 7. While MP2/TZVPP shows nonlinear curves for the populations of different clusters with varying temperature, the other electronic structure methods lead to a linear behavior for almost all clusters. Both BP electronic structure calculations populate the **s9** cluster far too much if compared to the MP2/TZVPP reference. Furthermore, the **s9** cluster population is still larger than the ring cluster population at high temperature for BP. This result might be comparable to the findings of Shields and Kirschner, who observed that some specific structures if calculated with DFT are not minimum structures.[25] Please note that comparing MP2 populations with large and small basis sets also shows differences.

4.4. Cooperativity. For both the $\mathbf{2_{opt}}$ set and the $\mathbf{2-4_{opt}}$ set we again list the accuracy and the QCE parameters in Table 5 together with the results of the pair energies applied to both optimal sets. The derivation of pair energies is described in refs 4 and 6. Comparing the a_{mf} values for both sets obtained with and without inclusion of cooperative effects, we find that a_{mf} increases by a factor larger than 2 for the calculations based on the pair energies. The difference in a_{mf} comparing the two sets is also present if we consider the QCE results obtained from the pair energies. In contrast to the behavior of the cooperative energies, the a_{mf} value for the pair energies (Table 5) is larger for the $\mathbf{2-4_{opt}}$ set than for the $\mathbf{2_{opt}}$ set, corresponding to the fact that cooperative effects in the spiro clusters with tetrahedrally coordinated water are more severe than in the $\mathbf{2_{opt}}$ cluster set.

The pair populations (not depicted) for both optimal sets show a decreasing dimer population from 86% to approximately 27% and an increase in the monomer population with increasing temperature to the extent of 27%, while the larger clusters are not significantly populated at all. Again these findings are comparable to the results obtained by Lee and Tuckerman.[24] The authors compared conditional correlation functions for hydrogen bonds in water as obtained from traditional molecular dynamics simulations with a nonpolarizable model to their first-principles simulations. While the functions decayed for all choices of coordination environment with the same speed in the traditional molecular dynamics simulations, first-principles simulations could show that the function for water with a 4-fold coordination decreases significantly slower, indicating that tetrahedrally coordinated water might possess extra stability.[24]

5. Discussion

Reconsidering the hydrogen bond, it is obvious that the strength of the particular hydrogen bond should be an objective. Weak and strong hydrogen bonds per particular water molecule are discussed in an asymmetric model.[26] From the point of view of static calculations we observe that some tetrahedrally coordinated water exhibits asymmetric coordination; i.e., individual hydrogen bonds are of different stabilities. However, if an average is taken over each of the tetrahedrally coordinated water molecules in an ensemble, this could wash out subtle effects, and as a result the four hydrogen bonds would be of similar strength. Interestingly, one of our clusters, namely, **s9**, shows an almost perfectly symmetric central water molecule (NBO[27,28] occupation numbers of each accepting σ^* orbital of the four hydrogen bonds, obtained with HF/SVP: 0.036, 0.036, 0.034, and 0.034), while the other spiro clusters deviate much more from perfect symmetry. For example, **s11** shows occupation numbers at the AADD water molecule of 0.035, 0.033, 0.024, and 0.041. It might be important that the **s9** cluster with the closest symmetric coordination plays such a dominant role;

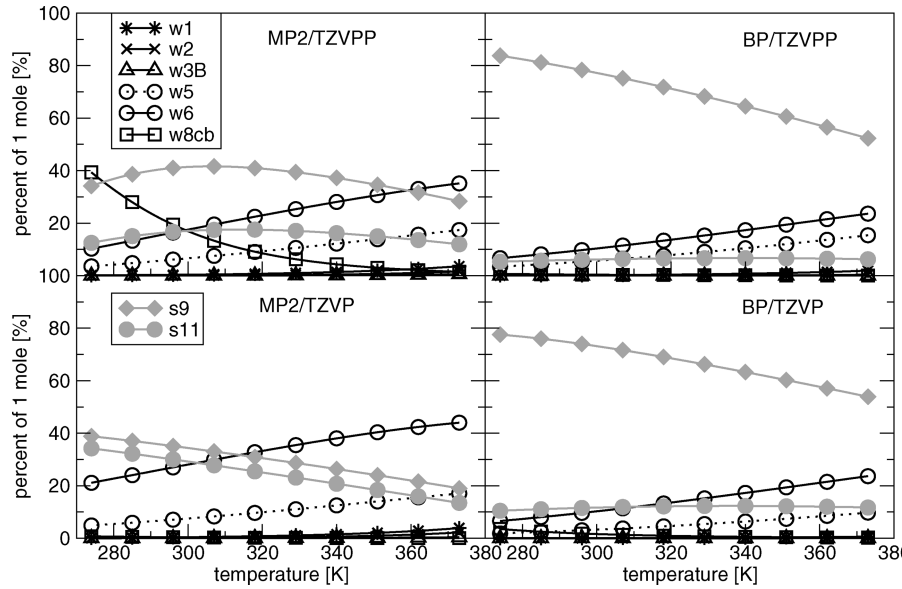


Figure 7. Populations at the liquid-phase temperature range for the **2**−**4**$_{opt}$ set with $a_{mf} = 0.130$ and $b_{xv} = 1.105$ at constant MP2/TZVPP frequencies.

Table 6. Fit for the Elimination of Different Clusters in the Temperature Range from 274 to 373 K[a]

	deleted?															
w3A	no	yes	yes	yes	**yes**	no	no	no	yes	yes	yes	no	no	no	no	yes
w3B	no	yes	yes	yes	**no**	yes	no	yes	yes	no	no	yes	no	no	yes	no
s7	no	yes	yes	no	**yes**	yes	yes	yes	no	yes	no	no	no	yes	no	no
s13	no	yes	no	yes	**yes**	yes	yes	no	no	no	yes	yes	yes	no	no	no
$\|\|\Delta V\|\|$	166	144	164	141	**137**	145	144	169	157	162	138	144	140	169	165	163

[a] The accuracy $\|\|\Delta V\|\|$ is given in the last line in dimensions of microliters. The electronic structure method is MP2/TZVPP.

see Figure 7. The highly populated **w6** cluster is perfectly symmetric as well with occupation numbers of 0.034, while the AD water molecules in the spiro clusters show slightly lower occupation numbers.

From static quantum chemical methods information about the hydrogen bond can be obtained. However, these calculations are based on fixed nuclei positions, and a hydrogen bond definition is difficult, because the wave function cannot be unequivocally distributed among the atoms.[29,30] Furthermore, as Lee and Tuckerman put it nicely, "Although reproducing these (binding) energies is important, small clusters do not represent the bulk...", these investigations neglect the condensed-phase environment completely. Nevertheless, interesting analyses such as energy decomposition to determine the origin of hydrogen bonding were carried out.[30,31]

The QCE method[3,4,6,13] applied in this work is based on static cluster calculations neglecting an explicit dynamical behavior as explained in the Introduction. It includes two parameters, a_{mf} and b_{xv}, accounting for the missing intercluster interaction and for excluded volume effects. A direct comparison of different electronic structure models is difficult, because of this parametric dependency but also because of the approximations made within the method.[4,6] However, if the parameters together with the input for the other partition functions are kept constant, the influence of different electronic structure methods can be tested. Despite these difficulties, the QCE method has the advantage that it is up to now the only method to incorporate electronic correlation and cooperative effects in a simple way for calculations of the condensed phase. Furthermore, the method in principle covers every temperature and pressure point in phase as long as the approximations made in the model are valid at that state. Due to the analytical form of the QCE partition functions, calculations of partition-function-dependent quantities are easily carried out. Another advantage of the QCE method is that it allows calculation of populations of clusters and thereby determination of which structural motif plays a role in the examined phases.

6. Conclusion

We investigated the liquid-phase isobars of water as obtained from the quantum cluster equilibrium method for different electronic structure methods, namely, MP2 and BP, at different basis sets. Two cluster sets were employed, one where no tetrahedrally coordinated water molecule was present (**2/3** set) and one where such motifs were included (**2**−**4** set). Both sets contain the ring motif, but the **2**−**4** set comprises interconnected ring structures of two equally sized rings. If we apply the **2/3** cluster set and allow the two adjusting parameters of the QCE model to vary, we find only slight differences in the isobars between DFT and MP2. Using a smaller basis set leads to worse agreement as compared to using the larger basis set for both methods. Next, we introduced optimal sets by eliminating underpopulated clusters (**2**$_{opt}$ and **2**−**4**$_{opt}$). This seems to be an important step in the QCE procedure, if high accuracy is desired,

because some clusters destabilize the system. One set contained only pure 2-fold coordinated water molecules ($\mathbf{2_{opt}}$). For the **2/3** and the $\mathbf{2_{opt}}$ cluster set we find that a ring consisting of six monomer units plays a dominant role over the whole temperature range, as observed in previous studies.[5,7] Adding the spiro clusters, we find a significant improvement of our calculated data with respect to experiment, especially at low temperature and concerning the slope of the curve. The QCE parameter accounting for the intercluster interactions decreases by adding tetrahedrally coordinated water molecules, indicating the improved description of the liquid phase in terms of intracluster energetics. Keeping all model parameters constant (at the optimized MP2/TZVPP values) and varying only the electronic structure energies, we find large basis set dependencies for the BP methods. Deviations for the different electronic structure methods from the MP2/TZVPP data obtained with the $\mathbf{2-4_{opt}}$ cluster set are now more pronounced at higher temperature.

Acknowledgment. This work was supported by the DFG, in particular by the ERA Chemistry Program and by the SPP-1191 Program. Computer time from RZ Leipzig, HLRS Stuttgart, and NIC Jülich are gratefully acknowledged.

Note Added after ASAP Publication. This paper was released ASAP on May 13, 2009, with errors in eq 4 and the following line. The correct version was posted on May 19, 2009.

Appendix

In Table 6 we list the results of all possible reduced combinations. We find a combination where the accuracy is even improved to 1.37×10^{-4} L (bold in the table). We will thus work with the set where **w3A**, **s7**, and **s13** are deleted.

References

(1) Kirchner, B.; Wennmohs, F.; Ye, S.; Neese, F. *Curr. Opin. Chem. Biol.* **2007**, *11*, 134–141.

(2) Spickermann, C.; Felder, T.; Schalley, C. A.; Kirchner, B. *Chem.—Eur. J.* **2008**, *14*, 1216–1227.

(3) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 367–372.

(4) Kirchner, B. *Phys. Rep.* **2007**, *440*, 1–111.

(5) Spickermann, C.; Lehmann, S. B. C.; Kirchner, B. *J. Chem. Phys.* **2008**, *128*, 244506.

(6) Kirchner, B. *J. Chem. Phys.* **2005**, *123*, 204116.

(7) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 373–384.

(8) Ludwig, R.; Weinhold, F. *J. Chem. Phys.* **1999**, *110*, 508–515.

(9) Ludwig, R.; Weinhold, F. *Z. Phys. Chem.* **2002**, *216*, 659–674.

(10) Ludwig, R.; Weinhold, F.; Farrar, T. C. *J. Chem. Phys.* **1995**, *103*, 3636–3642.

(11) Ludwig, R.; Weinhold, F.; Farrar, T. C. *J. Chem. Phys.* **1997**, *107*, 499–507.

(12) Ludwig, R. *ChemPhysChem* **2005**, *6*, 1376–1380.

(13) Borowski, P.; Jaroniec, J.; Janowski, T.; Woliński, K. *Mol. Phys.* **2003**, *101*, 1413–1421.

(14) Wendt, M. A.; Weinhold, F.; Farrar, T. C. *J. Chem. Phys.* **1998**, *109*, 5945–5947.

(15) Ludwig, R.; Behler, J.; Klink, B.; Weinhold, F. *Angew. Chem., Int. Ed.* **2002**, *41*, 3199–3202.

(16) Ludwig, R. *ChemPhysChem* **2007**, *8*, 938–943.

(17) Song, H.-J.; Xiao, H.-M.; Dong, H.-S.; Huang, Y.-G. *J. Mol. Struct.: THEOCHEM* **2006**, *767*, 67–73.

(18) Huber, H.; Dyson, A.; Kirchner, B. *Chem. Soc. Rev.* **1999**, *28*, 121–133.

(19) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(20) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(21) Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. *J. Comput. Chem.* **2002**, *23*, 895–910.

(22) Lemmon, E. W.; McLinden, M. O.; Friend, D. G. Thermophysical Properties of Fluid Systems. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* [Online]; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD. http://webbook.nist.gov (accessed 2005).

(23) Kirchner, B.; Spickermann, C. *PEACEMAKER*, V1.4 2004-2008; Institute of Physical and Theoretical Chemistry, University of Bonn: Bonn, Germany; Wilhelm-Ostwald Institute of Physical and Theoretical Chemistry, University of Leipzig: Leipzig, Germany, 2008.

(24) Lee, H.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.

(25) Shields, G. C.; Kirschner, K. N. *Synth. React. Inorg., Met.-Org., Nano-Met. Chem.* **2008**, *38*, 32–39.

(26) Soper, A. K. *J. Phys.: Condens. Matter* **2005**, *17*, 3273–3282.

(27) Weinhold, F. *Adv. Protein Chem.* **2006**, *72*, 121–155.

(28) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(29) Reiher, M.; Kirchner, B. *J. Phys. Chem. A* **2003**, *107*, 4141–4146.

(30) Thar, J.; Kirchner, B. *J. Phys. Chem. A* **2006**, *110*, 4229–4237.

(31) Morokuma, K. *Acc. Chem. Res.* **1977**, *10*, 294–300.

CT800310A

JCTC Journal of Chemical Theory and Computation

Quantum Cluster Equilibrium Theory Applied in Hydrogen Bond Number Studies of Water. 2. Icebergs in a Two-Dimensional Water Continuum?

S. B. C. Lehmann, C. Spickermann, and B. Kirchner*

Wilhelm-Ostwald Institute of Physical and Theoretical Chemistry, University of Leipzig, Linnéstrasse 2, D-04103 Leipzig, Germany

Received February 3, 2009

Abstract: With the aid of the quantum cluster equilibrium method, we calculate thermodynamic properties for a new water cluster set containing 2-fold and additional tetrahedrally hydrogen-bonded water molecules on the basis of accurate correlated electronic structure calculations. The addition of clusters with 4-fold coordinated water molecules leads to an improved thermodynamical description of the liquid phase in comparison to experimental values. The comparison of the obtained isobars from the pure 2-fold cluster set with the mixed cluster set shows improved results for the mixed set. Furthermore, the results of the liquid-phase entropy calculation compare excellently with experiment if the mixed cluster set is applied. The calculated populations allow us to determine hydrogen bond numbers, resulting in a temperature-dependent average hydrogen bond number. We observe a decreasing average hydrogen bond number of 2.77 at 274 K to 2.26 at 373 K and a dominance of 75% 2-fold hydrogen-bonded water molecules at room temperature for the mixed cluster set.

1. Introduction

For many decades scientists have been trying to reveal the complicated structure of liquid water to understand why it is such an important solvent. Already in 1892 Wilhelm Conrad Röntgen characterized the constitution of liquid water as an aggregate of two types of water; the first type of molecules he denoted as ice molecules.[1] Furthermore, he explained that both types of molecules are needed to predict the point of maximum density at 4 °C by the two opposing processes during cooling. These are on one hand a dilatation and on the other hand the common thermal contraction of liquids. According to that, the first type of molecules will change into the second type of molecules by heating the liquid and vice versa. Most later publications on the matter of liquid water highlight the need of tetrahedrally coordinated molecules for the occurrence of the point of maximum density as well. However, since the challenging paper of Wernet et al.,[2] the coordination number of water molecules in the liquid phase again has become a topic of heated debate in the literature. For example, Ball opens his essay on water with the following sentence: "No one really understands water." [3] Although the textbook opinion of the coordination number of water dissolved in itself was three to four, Wernet et al. found in their study indications for a coordination number of mainly two.[2] The authors applied X-ray absorption spectroscopy (which is sensitive to local hydrogen bond patterns) next to density functional theory calculations.[2] The emerging picture was a two-state pattern of the water hydrogen bond, namely, a "random soup" (hydrogen-bonded chains or rings) with "tiny icebergs" (tetrahedrally coordinated water).[3] Many papers on this subject (the number at the end of the year 2008 is approximately 270) followed, and we mention here only a selection. One of the critical papers was published by Head-Gordon and Johnson.[4] The authors carried out X-ray scattering experiments and inferred from those that asymmetry is inconsistent with their data, indicating fluctuations in the local molecular water environment.[4] An asymmetric water charge model for the site−site potential was investigated by Soper.[5] Soper found that neither the asymmetric model is correct nor the symmetric model is incorrect, but that X-ray and neutron diffraction data on

* Corresponding author phone: 493419736401; fax: 493419736399; e-mail: bkirchner@uni-leipzig.de.

10.1021/ct900189v CCC: $40.75
Published on Web 05/13/2009

water are rather insensitive to these details.[5] This was also discussed by Lee and Tuckerman. They concluded that overstructuring found in radial distribution functions (RDFs) calculated from *first-principles* simulations does not necessarily imply more rigid hydrogen bonds. Furthermore, the authors gave a warning toward deriving local structures of water from averaged quantities such as the RDFs.[6,7]

A remarkably outstanding paper describing a study in which traditional molecular dynamics simulations were applied in modified water was published recently.[8] Chatterjee, Debenedetti, Stillinger (who simulated, together with Rahman, water for the first time[9]), and Lynden-Bell investigated the effects of a water model that induces a 2-fold hydrogen bond only.[8] This was carried out in line with the previously studied idea to change the molecular dynamics potential parameters of Bergman and Lynden-Bell.[10] It was shown that a tetrahedrally coordinated hydrogen bond network is necessary to reproduce the water density anomaly. Head-Gordon and Rick also showed that classical molecular dynamics simulations lead to an erroneous description of the liquid phase of water, if a water model with an environment of two hydrogen bonds only is applied.[11] The origin of the density maximum in water was recently studied by Deeney and O'Leary.[12] The authors accounted for the density maximum in terms of opposing action of two independent physical processes,[12] one of these processes being a classical expansion/contraction effect and the other being identified as quantum zero-point-energy fluctuations. These effects counterbalance each other, resulting in a density maximum.[12]

The observations of ref 6 are in line with the finding of Ludwig, who could also show that only the tetrahedrally coordinated water structures are able to reproduce the density maximum of water.[13] On the basis of Hartree−Fock and density functional theory in the framework of the quantum cluster equilibrium (QCE) method, Ludwig discussed the importance of the tetrahedrally coordinated water versus a twice hydrogen-bonded water molecule. By inclusion of a tetrakaidecahedral $(H_2O)_{24}$ cluster, a triple point could be determined. Furthermore, Ludwig could show that the three-dimensional water clusters $(H_2O)_{13}$, $(H_2O)_{15}$, and $(H_2O)_{17}$ including tetrahedrally coordinated water molecules are necessary to mimic liquid-phase properties.[13] The voluminous water clusters are consistent with the oxygen−oxygen pair correlation function from X-ray diffraction experiments, and these clusters are able to reproduce the main features of the OH stretch region in the IR spectrum.[13]

In the present study we apply the quantum cluster equilibrium method to investigate hydrogen bond patterns in liquid water. We focus on the discussion of populations along the liquid temperature range, i.e., from 274 to 373 K.

This paper is structured as follows. First, a short methodological introduction to the QCE method as well as the calculation of the hydrogen bond numbers is given. Second, we compare the isobars of two cluster sets containing either 2-fold or 2-, 3-, and 4-fold hydrogen-bonded clusters. Next, cluster populations are considered. Thereafter, we show the improved results of the entropy calculations with rising temperature. Finally, we analyze the calculated temperature-dependent hydrogen bond numbers. The paper ends with the Discussion and Conclusion.

2. Methodology

The QCE theory and the computational details are described in the accompanying article (10.1021/ct800310a).[14] In the QCE approach modulated partition functions from static quantum chemical calculations are applied to obtain polynomials which are solved in a self-consistent fashion. The resulting populations are used to obtain bulk partition functions, and from those it is possible to calculate bulk thermodynamic properties.[15−17] To evaluate the importance of distinct cluster structures, we show monomer-normalized populations in this study as introduced in the previous paper (10.1021/ct800310a).

Cluster populations within our QCE program[18] usually refer to the total number of clusters. This number varies within a QCE calculation. If mainly large clusters are populated, more monomers are bound within these large clusters. Because the amount of monomers is fixed in the QCE calculation, there are fewer monomers to form the other clusters. This implies a reduction of the total number of clusters; i.e., if the total number of clusters decreases, the percentage population of each individual cluster increases. This is not useful when insight into the physical nature of a phase point is desired. For this purpose it is better to analyze populations which are related to the monomer reference system. Given that the total number of monomers is equal to 1 mol for all phase points, the percentage now provides an estimate of how many monomers of the total 1 mol are bound in a particular cluster and thereby reflects the physical composition of the phase point.

Furthermore, we analyze the average hydrogen bond number at each phase point as described in the following. For each cluster j and temperature the QCE calculation provides the *intercluster* interaction in terms of the mean field energy E_j^{inter}, depending on the cluster size and the calculated volume as well as the selected a_{mf} value:

$$E_j^{\text{inter}} = -i_j a_{\text{mf}} V^{-1} \tag{1}$$

This energy needs to be correlated to a hydrogen bond number (n^{hb}). Therefore, it will be divided by the cluster-specific intracluster interaction energy per hydrogen bond E_j^{hb}, the latter being obtained, e.g., by a natural bond orbital (NBO) analysis, leading to

$$n_{\text{inter},j}^{\text{hb}} = \frac{E_j^{\text{inter}}}{E_j^{\text{hb}}}, \quad \text{with} \quad E_j^{\text{hb}} = \frac{\Delta E_j^{\text{intra}}}{n_j^{\text{hb}}} \tag{2}$$

Alternatively, the mean field energy could be divided by an average binding energy per hydrogen bond E^{hb} given for each cluster set, similar to the energy criterion for hydrogen bonds in molecular dynamics simulations:

$$n_{\text{inter}}^{\text{hb}} = \frac{E_j^{\text{inter}}}{E^{\text{hb}}} \tag{3}$$

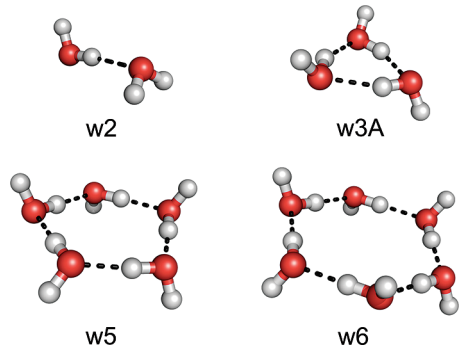


Figure 1. Ball-and-stick model of the 2-fold water cluster set (abbreviated as the $\mathbf{2_{opt}}$ set) as introduced in the previous paper (10.1021/ct800310a).[14]

The calculation of these additional hydrogen bonds arising from the mean field interaction (n^{hb}_{inter}) of a specific cluster has to be repeated for each cluster of the particular set. To derive the global average number of hydrogen bonds in the system, these cluster-specific additional hydrogen bonds have to be weighted by the monomer-normalized population of each cluster N_j, with $\sum_{j=1}^{\eta} N_j = 1$ and η being the total number of clusters. After the calculation of all hydrogen bonds and the weighting by their respective populations, they are added to yield the global average hydrogen bond number, leading to the final equation

$$\langle n^{hb} \rangle = \sum_{j=1}^{\eta} (n^{hb}_{inter,j} + n^{hb}_{j}) N_j \qquad (4)$$

It should be noted here that the above sketched scheme is not from first principles, because (a) an empiric parameter (a_{mf}) enters our model and (b) we distribute this energy not unequivocally. The optimized cluster sets employed in the present study are taken from the previous paper (10.1021/ct800310a).[14] In the present work we only show MP2/TZVPP electronic structure data. The $\mathbf{2-4_{opt}}$ set denotes the optimal cluster set containing clusters with tetrahedrally coordinated water molecules (see Figure 2), whereas the $\mathbf{2_{opt}}$ set only contains clusters with 2-fold hydrogen-bonded water molecules; see Figure 1.

The applied QCE parameters are listed in Table 1. A detailed discussion can be found in the accompanying paper (10.1021/ct800310a).[14]

3. Results

3.1. Comparing the Pure 2-Fold Water Set ($\mathbf{2_{opt}}$) to the Mixed Set ($\mathbf{2-4_{opt}}$). In the former paper (10.1021/ct800310a) we found an improvement with respect to the accuracy as well as a reduction of the mean field parameter a_{mf} especially at low temperature if tetrahedrally coordinated water molecules were included in the QCE calculations.[14] To further discuss this point, we show the resulting isobars in Figure 3.

We observe a definite improvement of the isobar due to the inclusion of additional spiro clusters. The slope of the experimental curve is reproduced better, and the previously discussed larger deviation at lower temperature is also reduced. This result clearly points at the importance of tetrahedrally coordinated water, which has to be applied in the calculation of the partition functions at low temperatures. It is noteworthy to remark that this result cannot be achieved by selecting the QCE parameters accordingly. Only the explicit treatment of the tetrahedrally coordinated water leads to improved results, which is a strong point in favor for the applicability of the QCE method and which shows that the QCE is sensitive to structural motifs of the condensed phase.

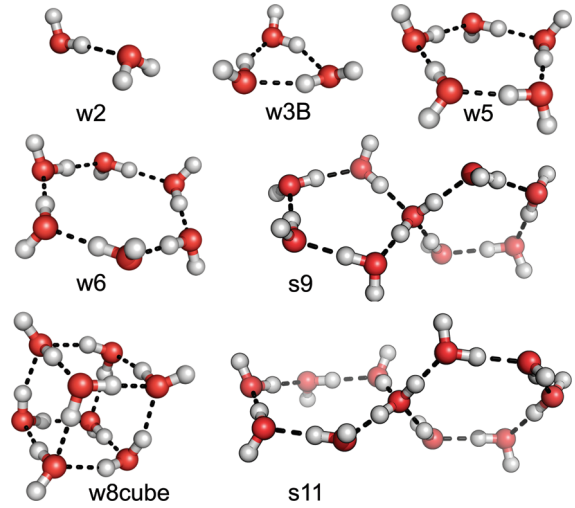


Figure 2. Ball-and-stick model of the cluster set containing tetrahedrally coordinated water molecules (abbreviated as the $\mathbf{2-4_{opt}}$ set) as introduced in the accompanying paper (10.1021/ct800310a).[14]

Table 1. QCE Parameters for the Sets without Spiro Clusters (**2/3** and $\mathbf{2_{opt}}$) and with the Spiro Clusters ($\mathbf{2-4_{opt}}$) at the Temperature Range from 274 to 373 K[a]

set	$\|\|\Delta V\|\|$	a_{mf}	b_{xv}
2/3	311.80	0.162	1.107
$\mathbf{2_{opt}}$	167.57	0.136	1.088
$\mathbf{2_{opt}}$ pair	405.63	0.343	1.068
$\mathbf{2-4_{opt}}$	137.07	0.130	1.105
$\mathbf{2-4_{opt}}$ pair	420.18	0.351	1.071

[a] $\|\|\Delta V\|\|$ (μL) gives the accuracy of the corresponding isobar, i.e., the root mean square deviation from the experimental values.

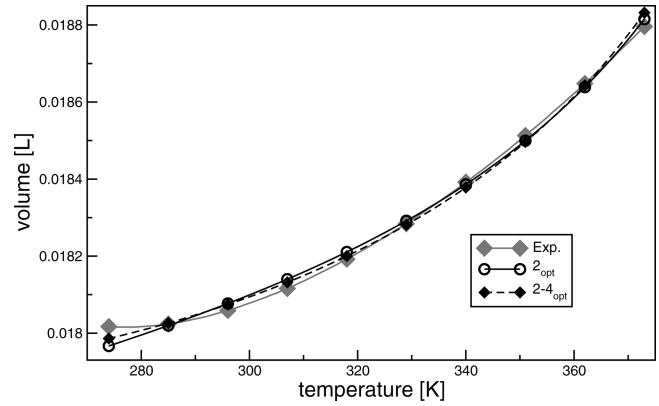


Figure 3. Calculated isobars for the $\mathbf{2_{opt}}$ (solid line and circles) and the $\mathbf{2-4_{opt}}$ (dashed line and tilted squares) cluster sets at the temperature range from 274 to 373 K with fitted parameters.

3.2. Populations of Clusters in the Bulk. According to QCE methodology, cluster populations are obtained at

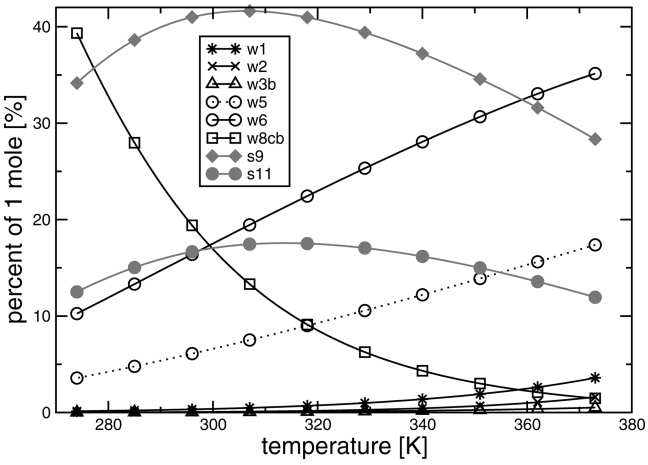


Figure 4. Monomer-normalized populations for the liquid-phase temperature range for the optimal **2−4$_{opt}$** set.

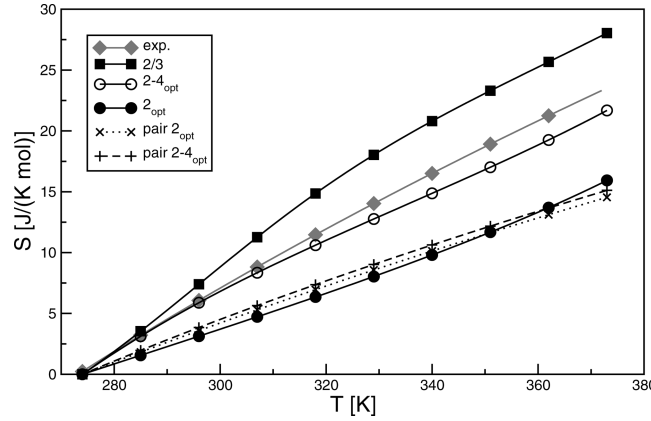


Figure 5. Entropy as calculated according to ref 17. The reference point for the calculations and the experimental data is set to $S^{(273.16K)} = 0$.

every phase point of a QCE calculation, thereby providing insight into the composition of the bulk. This constitutes a link between isolated clusters obtained from static quantum chemical calculations and the bulk, as mentioned in the Introduction of the accompanying paper (10.1021/ct800310a).[14] In Figure 4 we depict the monomer-normalized populations; these populations show to what extent a particular cluster is populated with respect to the constant total monomer number of 1 mol.

There are three regions in which different clusters are mainly populated. At very low temperature the cagelike **w8cube** cluster shows a population of 40%, between 280 K and approximately 360 K the **s9** spiro cluster plays the important role with a maximum population of approximately 42% at 306 K, and at high temperature (>362 K) the **w6** cluster is most highly populated with 35%. While the populations of all clusters with only 2-fold coordination grow with increasing temperature, the spiro clusters show each a maximum. These maxima are approximately at 310 K for **s9** as well as for the **s11** cluster. The cagelike **w8cube** cluster, which seems to be an important motif of the low-temperature region, decreases much faster than the ring clusters increase. Thus, the QCE model predicts a significant temperature dependence of the liquid-phase coordination pattern, which has also been observed in recent experiments.

The importance of the investigated temperature is also emphasized in the paper of Head-Gordon and Johnson.[4] With the aid of temperature-dependent experiments, the authors draw the conclusion of a mainly tetrahedrally coordinated water structure. These conclusions are based on three temperature measurements, i.e., at 1 °C (274 K), 25 °C (298 K), and 77 °C (360 K). As can be seen from Figure 4, **s9** shows a maximum at approximately 306 K and the ring structures are just starting to show a growing population. Thus, the QCE results imply that the complete picture might be more complicated and that a simple linear temperature dependence might not apply. According to these QCE calculations, it would seem to be helpful to add more temperature-dependent measurements, for example, at 283 K (10 °C) and at 320 K (47 °C). The temperature-dependent behavior was also observed by Chatterjee et al. in the temperature-dependent spatial distribution functions plot for their modified model with a water H−O−H angle of 90° and 100°. These two models acquire a waterlike (tetrahedral) structure upon cooling.[8]

3.3. Entropy. One of the strong points of the QCE method is the possibility to calculate different thermodynamic quantities once the partition functions are known. Please note that once the QCE parameters are set they are not changed anymore; i.e., the outcome of the calculations does not depend on the selection procedure of the QCE parameters, but on how well the model-inherent approximations work for the particular properties. In a previous study we employed QCE calculations to derive entropies of the liquid phase as well as the liquid−vapor phase transition.[17] The neglect of cooperativity led to large errors in the obtained entropy values. In contrast, a correct treatment of the intracluster many-body interaction yielded liquid-phase entropies and phase transition entropies in very good agreement with the experimental reference.[17]

In Figure 5 we show the entropy plotted against temperature for our different sets. As can be observed, the original **2/3** set lies above the experimental values; i.e., the slope of the curve is too large. All other methods underestimate the experimental reference. To obtain a closer agreement with experiment, tetrahedrally coordinated water molecules are necessary as reflected in the excellent agreement of the **2−4$_{opt}$** entropies with the experimental values; see Figure 5, black curve with open circles. To facilitate the relevance of the **2−4$_{opt}$** cluster set, it should be mentioned that the vaporization entropy $\Delta_{vap}S$ (not depicted), calculated with this set, also shows a more accurate value of 109.42 J/(K mol) (exp.: 109.06 J/(K mol)).[19] Applying 2-fold hydrogen-bonded clusters alone leads to entropies as inaccurate as the ones depending on pair energies only; see Figure 5, black curve with closed circles and dashed and dotted curves. This indicates that from the entropical point of view the **2$_{opt}$** cluster set and the sets applying the pair energies underestimate the entropy.

3.4. Hydrogen Bond Numbers. In the last section, we showed monomer-normalized populations for calculations with the **2−4$_{opt}$** set. However, the mean field QCE parameter prevents us from giving a quantitative statement of the average hydrogen bond for water molecules that could be

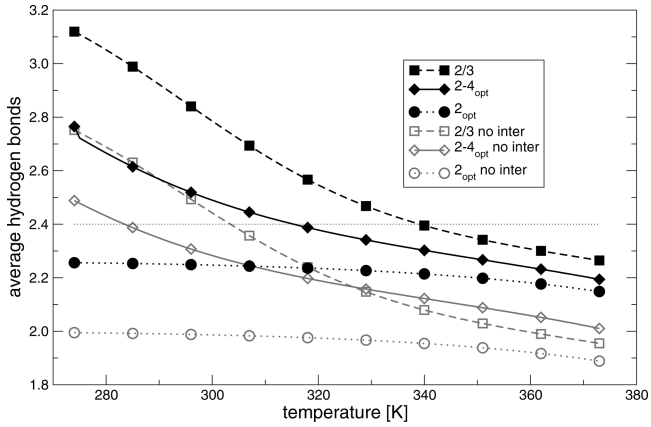


Figure 6. Average number of hydrogen bonds at different phase points with intercluster interaction recalculated as described in section 2.

expected from our calculations, since the number of hydrogen bonds accounted for by the mean field energy is only accessible in an indirect way (see section 2). While in the $\mathbf{2}_{\mathbf{opt}}$ clusters the maximum average hydrogen bond number that occurs is 2, the average number is 3.0 for the **w8cube** cluster and 2.22 for the **s9** cluster in the $\mathbf{2-4}_{\mathbf{opt}}$ set. To estimate the number of hydrogen bonds according to section 2, we have to evaluate the mean field energy in terms of the average hydrogen bond energy per cluster. The procedure to obtain the total average hydrogen bond number at different phase points is described in section 2.

In Figure 6 the average hydrogen bond numbers are depicted against temperature. We observe that for all sets these numbers decrease with increasing temperature; i.e., none of the sets provides hydrogen bond numbers that behave identically at every phase point, thus showing the temperature dependency of the hydrogen bond number as discussed in the earlier section. Considering the average hydrogen bonds in Figure 6, we find a monotonically decreasing behavior. This monotonical behavior could not be deduced directly from the populations. The QCE-predicted hydrogen bond numbers for set $\mathbf{2-4}_{\mathbf{opt}}$, set **2/3**, and set $\mathbf{2}_{\mathbf{opt}}$ do not reflect the picture of a mainly tetrahedrally hydrogen-bonded water. Furthermore, it is apparent from Figure 6 that the mean field term corrects the intercluster interaction in different ways. At lower temperature a higher hydrogen bond amount is added in all sets. This is the reason why the uncorrected curves ($n^{\mathrm{hb}}_{\mathrm{inter}} = 0$) at higher temperature (gray curves in Figure 6) show an order different from that of the curves including $n^{\mathrm{hb}}_{\mathrm{inter}}$ from the intercluster interactions. In general, the curves are a little bit more spread out at low temperature than at high temperature. At high temperature (373 K) the curves including $n^{\mathrm{hb}}_{\mathrm{inter}}$ exhibit values between 2.14 ($\mathbf{2}_{\mathbf{opt}}$) and 2.26 ($\mathbf{2-4}_{\mathbf{opt}}$), which indicates a high probability for 2-fold hydrogen-bonded water molecules at this phase point (87−93%) in the calculated cluster phase.

The maximum hydrogen bond number (3.12) is obtained for the **2/3** set at lowest temperature (274 K). At the same temperature the value for the $\mathbf{2}_{\mathbf{opt}}$ set amounts to 2.26 and for the $\mathbf{2-4}_{\mathbf{opt}}$ set to 2.77. At this stage it is necessary to keep in mind that the isobars and entropies were most accurate for the $\mathbf{2-4}_{\mathbf{opt}}$ set; i.e., the obtained data from this set should be taken as the most reliable results. This reasoning is based on the assumption that the most accurate thermodynamics also yield the most precise structural information. The hydrogen bond numbers indicate (if we neglect the fact that 3-fold hydrogen bonding occurs) 38% tetrahedrally coordinated water molecules and 62% 2-fold hydrogen bonds for the $\mathbf{2-4}_{\mathbf{opt}}$ set, 13% tetrahedrally coordinated water and 87% 2-fold hydrogen-bonded water molecules for the $\mathbf{2}_{\mathbf{opt}}$ set, and 56% tetrahedrally coordinated water molecules and 44% 2-fold hydrogen bonds for the **2/3** set at 274 K. A ratio of 80% tetrahedrally to 20% 2-fold coordinated water would be represented by a hydrogen bond number of 3.6. This clearly shows that our results are to a greater extent in accordance with the findings of Wernet et al.;[2] i.e., we obtain the closest agreement to the experimental thermodynamics (e.g., see Figure 3 in section 3.1 and Figure 5 in section 4) with mainly 2-fold coordinated water molecules (75% at 298 K with the $\mathbf{2-4}_{\mathbf{opt}}$ set). It should be noted here that these numbers are based on the approximate evaluation of the mean field energy (see section 2) and thus should be taken as semiquantitative. In traditional molecular dynamics simulations values from 3.1 to 3.3 hydrogen bonds at room temperature are discussed.[11] The hydrogen bond number of 2.4 could be attributed to a ratio of 80% 2-fold hydrogen bonds to 20% tetrahedrally coordinated water molecules (see the dotted horizontal line in Figure 6) with the hypothetical assumption of these two possibilities only. From approximately 315 K the total hydrogen bond number of the $\mathbf{2-4}_{\mathrm{opt}}$ set drops below 2.4; see the dotted line in Figure 6. However, the limited size of the clusters present in the $\mathbf{2-4}_{\mathbf{opt}}$ set makes it difficult to realize a distribution of 80% 4-fold to 20% 2-fold coordination at all, because in these medium-sized clusters the number of 2-fold coordinated molecules will always be larger than the number of 4-fold coordinated molecules due to surface effects. Nevertheless, these QCE results demonstrate that the coordination number distribution present in the $\mathbf{2-4}_{\mathrm{opt}}$ set is consistent with experimental thermodynamics, at least concerning densities and entropies.

Applying the binding energy criterion of the dimer as mentioned earlier in section 2 (not depicted) instead of the cluster-specific binding energy for a hydrogen bond results in a 4.9% higher hydrogen bond number for the $\mathbf{2-4}_{\mathbf{opt}}$ set, while the **2/3** set shows a difference of 5.9% and the $\mathbf{2}_{\mathbf{opt}}$ set shows the largest difference with an increase of 6.2%.

If we compare the curves from the cooperative energies with those of the pair energies (not depicted), we find that the pair curves of both sets obviously exhibit much lower coordination numbers. For the uncorrected case ($n^{\mathrm{hb}}_{\mathrm{inter}} = 0$) both sets exhibit an n^{hb} of below 1, while the corrected sets show a hydrogen bond number increased to an amount of more than 1, $\langle n^{\mathrm{hb}} \rangle$.

4. Discussion

To clarify several indistinct issues, it seems to be important to recapitulate different aspects of theoretical investigations, and additionally we want to reconsider some structural aspects of liquid water.

(1) Water has a coordination number (as opposed to the hydrogen bond number examined in the present study) that is given by the number of surrounding water molecules or the first integrated X-ray oxygen−oxygen radial distribution function $g_{oo}(r)$ peak. As Soper explains, the question of this coordination number is distinct from that concerning the number of hydrogen bonds.[5] Some scientists infer that water molecules mainly contain two hydrogen bonds, and some believe that water largely shows a tetrahedral hydrogen bond pattern in the liquid phase.[3]

(2) From the theoretical perspective the water structure was mainly investigated by traditional molecular dynamics (MD) simulations.[8,20] These methods include a model-inherent dynamical description and large samples (Soper used 1800 water molecules[5]). Nevertheless, the methods are mostly applied with fixed charges (even if these are distributed asymmetrically) and with the pairwise additivity approximation as well as the neglect of nuclear quantum effects.[21] Many suggestions for polarizable water models appeared in the literature.[22] The quality of parametrization varies from system to system and from quantity to quantity, raising the question of transferability.[23] Despite these problems, it is possible to reproduce such important quantities as the density maximum with traditional MD simulations.[8]

(3) First-principles simulations, for instance, in the framework of Car−Parrinello[24] molecular dynamics simulations, circumvent the approximation of traditional molecular dynamics simulations. However, within these methods it is only possible to treat a small sample (∼1000 molecules; see ref 25) with a short simulation time (<100 ps scale). Due to large computational costs, the calculations are mainly carried out with density functional theory (neglecting dispersion) and relatively small basis sets. Hybrid functionals[26] and correction schemes for dispersion[27−29] were successfully tested. A very valuable discussion of different effects is given by Lee and Tuckerman.[7] One of the major advantages of first-principles simulations is that the electronic structure can be analyzed on the fly.[30] For example, it is possible to calculate local dipole moments[31,32] or charges[33,34] in the liquid phase.

(4) Despite the fact that only a few clusters in relation to MD simulations are applied within the QCE model, the cluster-inherent parameters as geometries and frequencies depend on accurate electronic structure models. Furthermore, the QCE method employs no effective potentials in the description of intermolecular interactions contrary to traditional molecular dynamics simulations, which makes it more sensitive to questions of hydrogen-bonding patterns. While it is difficult to employ an unambiguous criterion for hydrogen bonding in traditional molecular dynamics simulations, a less crude analysis of the clusters calculated with accurate electronic structure methods gives rise to the question of how many "hydrogen bonds" are inherent in a specific cluster. From those clusters further calculations such as the calculation of the populations or isobars were carried out, finally leading to the prediction of a hydrogen bond number in the temperature range of the liquid phase of water, still on the basis of the accurate electronic structure calculations. The QCE model is not able to describe the dynamics of hydrogen bond formation and breaking, but nonetheless, a strong point of the model is that temperature-dependent properties, e.g., the $\langle n^{hb} \rangle$ or the population of a distinct motif, of the investigated systems can be shown easily. Moreover, the QCE model can be used as a tool to trace deviations from the experiment due to certain characteristics such as geometries or electronic structure methods.

5. Conclusion

The extended water cluster set ($\mathbf{2}-\mathbf{4}_{opt}$), containing tetrahedrally hydrogen-bonded clusters, leads to an enhanced description of the liquid phase of water in the frame of the QCE theory. This result shows once more that the 4-fold coordination is necessary to obtain an accurate description of liquid water, e.g., in the case of the improved results in liquid-phase entropies. The analysis of the hydrogen bond patterns within our model clearly shows that, although the tetrahedrally coordinated water molecules have to be included in the cluster set to obtain an accurate physical behavior, the calculated cluster phase exhibits an average hydrogen bond number below 3 at room temperature. Furthermore, the application of the dimer hydrogen bond interaction energy alone as a criterion for hydrogen bonding leads to an erroneous because overstructured picture of the liquid phase. It might be inferred that traditional MD compensates the too low hydrogen bond energy of approximately 20 kJ/mol with an overstructuring (too many tetrahedrally coordinated water molecules) of the hydrogen bond network. Considering the averaged cooperative interaction energies of the higher populated clusters which dominate the QCE cluster phase, a much higher value (between 28 and 31 kJ/mol) is obtained. Thus, at equal total energy in the system the higher energy per hydrogen bond leads to a sparser hydrogen-bonded water network. This issue together with the poor performance of pairwise additive interaction energies applied in QCE calculations reconfirms the importance of cooperative effects in liquid water.[15]

Wernet et al. reinvestigated a conventional wisdom of water coordination with their set of experiments.[2] One of the benefits or side effects of these experiments lies in the improvement of other experiments, models and methods. At present no method or theory can claim to describe all features of water correctly. Therefore, the QCE is a helpful and necessary tool toward understanding this important task. It certainly provides us also with a link between isolated clusters and the condensed phase.

To gain additional insight into the local structure of liquid water, further studies in the frame of the QCE theory and thus on the basis of accurate electronic structure methods are necessary. Moreover, it is still a mandatory task to understand how the water molecules interact with each other on the molecular level.

Acknowledgment. This work was supported by the DFG, in particular by the ERA Chemistry Program and by the SPP-1191 Program. Computer time from RZ Leipzig, HLRS Stuttgart, and NIC Jülich is gratefully acknowledged.

References

(1) Röntgen, W. C. *Ann. Phys.* **1892**, *281*, 91–97.

(2) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Naslund, L.-A.; Hirsch, T. K.; Ojamae, L.; Glatzel, P.; Pettersson, L. G. M.; Nielsen, A. *Science* **2004**, *304*, 995–999.

(3) Ball, P. *Science* **2008**, *452*, 291–292.

(4) Head-Gordon, T.; Johnson, M. E. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7973–7977.

(5) Soper, A. K. *J. Phys.: Condens. Matter* **2005**, *17*, 3273–3282.

(6) Lee, H.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154507.

(7) Lee, H.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.

(8) Chatterjee, S.; Debenedetti, P. G.; Stillinger, F. H.; Lynden-Bell, R. M. *J. Chem. Phys.* **2008**, *128*, 124511.

(9) Rahman, A.; Stillinger, F. H. *J. Chem. Phys.* **1971**, *55*, 3336–3359.

(10) Bergman, D.; Lynden-Bell, R. M. *Mol. Phys.* **2001**, *99*, 1011–1021.

(11) Head-Gordon, T.; Rick, S. W. *Phys. Chem. Chem. Phys.* **2007**, *9*, 83–91.

(12) Deeney, F. A.; O'Leary, J. P. *Phys. Lett. A* **2008**, *372*, 1551–1554.

(13) Ludwig, R. *ChemPhysChem* **2007**, *8*, 938–943.

(14) Lehmann, S. B. C.; Spickermann, C.; Kirchner, B. *J. Chem. Theor. Comput.* **2009**, *5*, xxxx−xxxx (10.1021/ct800310a).

(15) Kirchner, B. *J. Chem. Phys.* **2005**, *123*, 204116.

(16) Kirchner, B. *Phys. Rep.* **2007**, *440*, 1–111.

(17) Spickermann, C.; Lehmann, S. B. C.; Kirchner, B. *J. Chem. Phys.* **2008**, *128*, 244506.

(18) Kirchner, B.; Spickermann, C. *PEACEMAKER*, V1.4 2004-2008; Institute of Physical and Theoretical Chemistry, University of Bonn: Bonn, Germany; Wilhelm-Ostwald Institute of Physical and Theoretical Chemistry, University of Leipzig: Leipzig, Germany, 2008.

(19) Lemmon., E. W.; McLinden, M. O.; Friend, D. G. Thermophysical Properties of Fluid Systems. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* [Online]; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD. http://webbook.nist.gov (accessed 2005).

(20) Dang, L. X.; Pettitt, B. M. *J. Chem. Phys.* **1987**, *91*, 3349–3354.

(21) Guillot, B.; Guissani, Y. *J. Chem. Phys.* **1998**, *108*, 10162–10174.

(22) Halgren, T. A.; Damm, W. *Curr. Opin. Biol.* **2001**, *11*, 236–242.

(23) Brodsky, A. *Chem. Phys. Lett.* **1996**, *261*, 563–568.

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

(25) Hutter, J. Private communication, 2008.

(26) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I. F. W.; Mundy, C. J. *J. Phys. Chem. B* **2006**, *110*, 3685–3691.

(27) Lilienfeld, O. A.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(28) Lilienfeld, O. A.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 014113.

(29) Lin, I.-C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Röthlisberger, U. *J. Phys. Chem. B* **2009**, *113*, 1127–1131.

(30) Iftimie, R.; Tuckerman, M. E. *J. Chem. Phys.* **2005**, *122*, 214508.

(31) Silvestrelli, P. L.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 3572–3580.

(32) Thar, J.; Reckien, W.; Kirchner, B. Car−Parrinello Molecular Dynamics Simulations and Biological Systems. In *Atomistic Approaches in Modern Biology*; Reiher, M., Ed.; Topics in Current Chemistry, Vol. 268; Springer: New York, 2007.

(33) Thar, J.; Zahn, S.; Kirchner, B. *J. Phys. Chem. B* **2008**, *112*, 1456–1464.

(34) Kirchner, B.; Hutter, J. *J. Chem. Phys.* **2004**, *121*, 5133–5142.

CT900189V

JCTC Journal of Chemical Theory and Computation

Unraveling the Catalytic Pathway of Metalloenzyme Farnesyltransferase through QM/MM Computation

Ming-Hsun Ho,[†] Marco De Vivo,[†,‡] Matteo Dal Peraro,[§] and Michael L. Klein*,[†]

Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104-6323, Department of Drug Discovery and Development, Italian Institute of Technology, Via Morego 30, I-16163 Genova, Italy, and Laboratory for Biomolecular Modeling, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland

Received November 4, 2008

Abstract: The protein farnesyltransferase (FTase) is a Zn^{2+}-metalloenzyme that catalyzes the farnesylation reaction, i.e., the transfer of the 15-carbon atom farnesyl group from farnesyl diphosphate (FPP) to a specific cysteine of protein substrates. Oncogenic Ras proteins, which are among the FTase substrates, are observed in about 20−30% of human cancer cells. Thus, FTase represents a target for anticancer drug design. Herein, we present a classical force-field-based and quantum mechanics/molecular mechanics (QM/MM) computational study of the FTase reaction mechanism. Our findings offer a detailed picture of the FTase catalytic pathway, describing structural features and the energetics of its saddle points. A moderate dissociation of the diphosphate group from the FPP is observed during the nucleophilic attack of the zinc-bound thiolate. At the transition state, a resonance structure is observed, which indicates the formation of a metastable carbocation. However, no stable intermediate is found along the reaction pathway. Thus, the reaction occurs via an associative mechanism with dissociative character, in agreement with the mechanism proposed by Fierke et al. (*Biochemistry* **2000**, *39*, 2593−2602 and *Biochemistry* **2003**, *42*, 9741−9748). Moreover, a fluorine-substituted FPP analogue (CF_3-FPP) is used to investigate the inhibitory effect of fluorine, which in turn provides additional agreement with experimental data.

Introduction

The protein farnesyltransferase (FTase), a Zn^{2+}-metalloenzyme, catalyzes the transfer of the 15-carbon farnesyl group from the farnesyl diphosphate (FPP) to acceptor proteins that contain the so-called "CaaX" motif at the C-terminus[1−3] (where C is the cysteine residue that is farnesylated, a is generally an aliphatic amino acid, and X is the terminal residue, which can be alanine, cysteine, serine, methionine, or glutamine[4−7]). FTase activity is crucial in signal transduction pathways such as proliferation and apoptosis of cells.[8,9] In fact, the Ras superfamily and small GTPases including Ras, Rho, and Rab are important examples of proteins that are activated by FTase function. Nowadays, FTase represents one of the promising targets for anticancer drug design,[10,11] being involved in the activation of oncogene proteins such as mutated Ras, which are related to the development of ∼20−30% of human cancers.[12,13]

FTase is a heterodimer, which consists of a 48 kDa α subunit and a 46 kDa β subunit.[3,14] Crystallographic and kinetic studies have suggested a two-step mechanism for substrates binding to FTase[15−17] whereby, initially, FPP binds to the hydrophobic cavity in the β subunit, and then the CaaX peptide substrate binds to form a ternary complex FTase/FPP/CaaX. One Zn^{2+} ion is accommodated in the

* Corresponding author. E-mail: klein@lrsm.upenn.edu. Phone: 215-898-8571. Fax: 215-898-5425.

[†] University of Pennsylvania.

[‡] Italian Institute of Technology.

[§] Ecole Polytechnique Fédérale de Lausanne.

10.1021/ct8004722 CCC: $40.75
Published on Web 05/21/2009

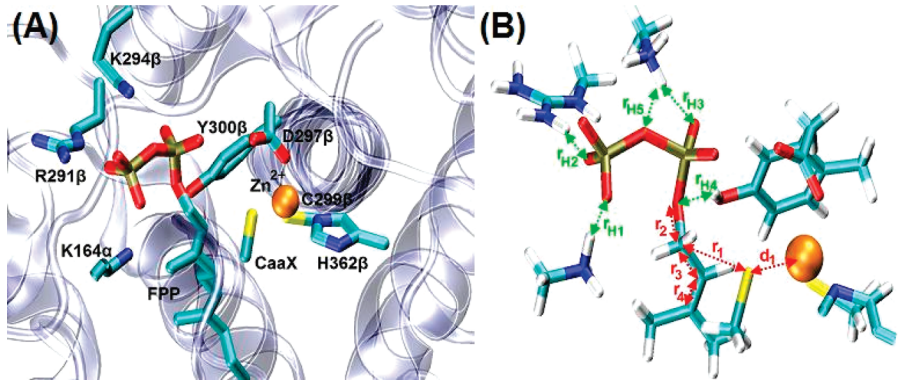


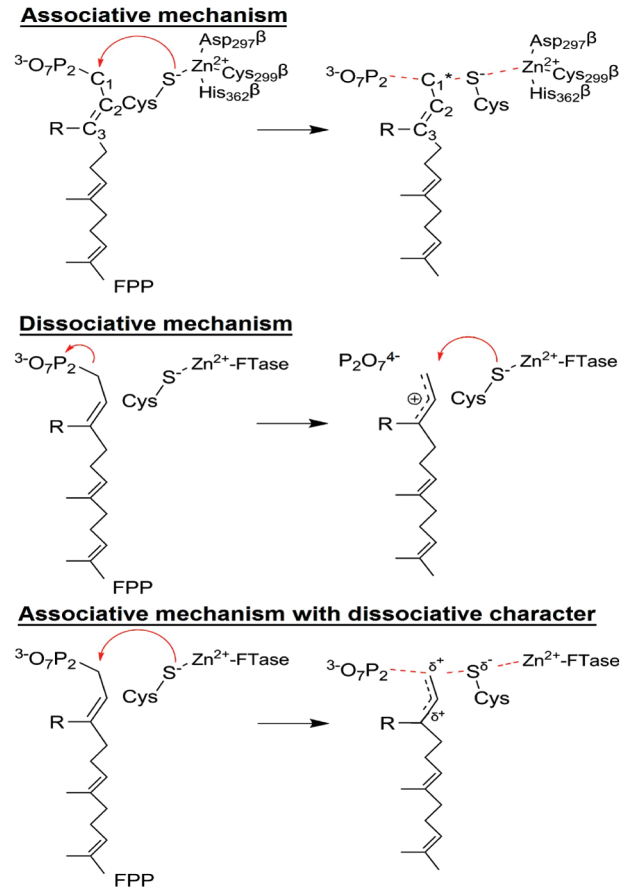
Figure 1. (A, Left Panel) A representative structure of the active site in the FTase/FPP/KCVIM ternary complex is shown. (B, Right Panel) the detailed conformation in the active site. The reaction coordinate is indicated by r_1, which is the distance between sulfur anion S^- and C_1 carbon on FPP.

FTase catalytic site, and its presence is essential for efficient enzymatic activity.[17,18] This metal ion is tetracoordinated to Asp 297β, Cys 299β, His 362β, and the CaaX cysteine residue (Figure 1A). Recent experimental and computational results indicate that the CaaX cysteine is a thiolate in the bound state at physiological pH.[19−21]

The FTase catalytic cycle involves two major steps: (1) the *physical step*, which consists of a conformational rearrangement needed to bring the FPP close to the nucleophilic thiolate in the ternary FTase/FPP/CaaX complex (the product of this step will be referred to as the *active* form hereafter), and (2) the *chemical step*, which is the transfer of the farnesyl group from the FPP to the thiolate.[22] In particular, in the physical step the large separation between the reactive atoms C_1 of the FPP substrate and the sulfur atom of the thiolate must be spanned for the enzymatic reaction to occur. Indeed, the large value reported for this distance in the crystal structures[5,23] ($r_1 = 7$ Å, Figure 1B) has led to the formulation of the so-called "distances paradox"[2] for which two hypotheses have been suggested: (1) Rotation of the first two isoprenoid subunits of the FPP, so as to bring the reactive C_1 atom closer to the thiolate,[17,22,24] or (2) dissociation of the thiolate from the Zn^{2+} ion and its subsequent approach to the FPP molecule.[25] Interestingly, a recent computational study has reported that the energy barrier for the rotation of the first two isoprene subunits of the FPP is lower than that of the dissociation of the thiolate from the Zn^{2+} ion coordination shell, thus favoring the first hypothesis.[26]

Three distinct hypotheses have been proposed concerning the chemical step, all of which are based on various experimental findings[2] (Scheme 1). The first hypothesis suggests an S_N1-like, "dissociative" mechanism, in which a stable carbocation is formed along the reaction pathway. Importantly, kinetic studies[27−29] have shown a significant decrease of the reaction rate when different fluoromethyl FPP analogues are used as ligands. This would be caused by the destabilization of the carbocation intermediate, thus indicating an S_N1-like mechanism for enzymatic activity. The second hypothesis suggests an "associative" S_N2-like mechanism. This is supported by stereochemical studies that have shown inversion of configuration during farnesylation,[30] as well as by an observation of a α-secondary kinetic isotope

Scheme 1. Three Distinct Reaction Mechanisms for the Farnesylation Reaction Have Been Proposed[a]



[a] (Upper) Associative mechanism. (Middle) Dissociative mechanism. (Lower) Associative mechanism with dissociative character. R = CH_3 in the FTase/CH_3-FPP/CaaX system, while R = CF_3 in the FTase/CF_3-FPP/CaaX system.

effect that has shown a value near unity.[31] Additional evidence for an associative S_N2-like mechanism comes from a metal-substitution experiment: a 6-fold decrease in reaction rate has been measured when the zinc ion has been substituted by a cadmium ion in the FTase complex.[29] In fact, cadmium ions form stronger metal−ligand interactions with thiolate than zinc ions.[32] As a result, this causes a decreased propensity for dissociation of the ligand (i.e., the

nucleophile) from the metal, which would lead to an associative interaction with FPP during the catalysis. Therefore, the decrease of the reaction rate in the Cd−FTase complex suggests an associative mechanism. Finally, a third mechanism has been proposed by Fierke et al. to explain the observation of both the nucleophilic and electrophilic character[20,24,29,33] of the reaction under study. In the Fierke model, the enzymatic mechanism is proposed to have an associative transition state (TS), with a dissociative character, which suggests a hybrid of the first two mechanisms; the main characteristic of this pathway is a modest dissociation of the FPP diphosphate group during the early phase of the nucleophilic attack of the zinc-bound thiolate on the FPP molecule.

In view of the above scenario, a more detailed picture of the reaction mechanism and its transition state would improve our understanding of FTase enzymatic activity. Toward this aim, we focus on the chemical step of the catalytic cycle and present a computational study of the farnesylation mechanism. Our computations employ classical molecular dynamics (MD) and ab initio Car−Parrinello[34] (CP) QM/MM calculations. Based on the crystallographic structures, two different systems have been used in this work: a complex that includes the natural FPP substrate, and a second complex that includes the trifluoromethyl-sutstituted FPP analogue, CF_3-FPP. Anticipating our results, we will see that, overall, our findings fit the experimental data rather well and thus provide additional insights into the nature of the farnesylation reaction mechanism and its inhibition by fluoromethyl-FPP analogues.

Methodology

Structural Models. A model of the FTase/FPP/CaaX ternary complex was generated based on the X-ray structure of Long et al.[22] (PDB entry code: 1KZP, 2.1 Å resolution). This farnesylated product contains the dephosphorylated FPP molecule, the CaaX peptide formed by the KCVIM sequence of amino-acids, and the FTase protein. One FPP molecule was used to replace the dephosphorylated FPP. The FPP molecule was added to the structural model in order to restore the initial state of the ternary complex. In particular, the diphosphate group was rotated to have a linear configuration of $O_1-C_1-S^-$ atoms (Figure 1B), while the isoprenoid chain matched the product form. The final model is consistent with the active form proposed by Long et al.[22]

In order to study the electron-withdrawing effect of fluoromethyl-FPP analogues, we replaced the CH_3 group on the first isoprenoid group with a CF_3 group in the model system. This CF_3-FPP analogue represents the ligand used by Dolence et al. for kinetic experiments.[28]

Molecular Dynamics. The FTase/FPP/CaaX ternary complex was immersed in a rectangular box of TIP3P waters (ca. 90 000 atoms in total). Classical MD was used to equilibrate the system and provide a suitable system for the subsequent QM/MM calculations. The AMBER force field (ff99)[35] was adopted for all standard residues, while RESP charges[36] were used for the Zn^{2+} ion, its ligands, and the FPP molecule. In order to simulate the FPP substrate, we adopted the parameters from Cui et al. for the isoprenoid part of the FPP residue.[37] After the initial setup, a 10 ns MD trajectory was performed with the NAMD package.[38] The Zn^{2+} tetracoordination configuration was restrained. Full details of the set up procedure are reported in the Supporting Information. The system reached convergence after the first 4 ns of dynamics (see Supporting Information). The relevant distances in the complex and in particular the conformation of the FPP substrate did not significantly change during the multins time scale of the MD trajectory. The average distance separating the carbon C_1 and sulfur anion was 3.50 ± 0.20 Å. A representative snapshot, chosen from the equilibrated part of the MD trajectory, was used for the following QM/MM investigation.

QM/MM Dynamical Studies. The enzyme-catalyzed reaction was investigated using the Car−Parrinello (CP) MD version of the quantum mechanical (QM)/molecular mechanics (MM) method,[39] which has been proven to be an excellent tool in investigating the reactivity of solvated biological systems, including metalloenzymes.[40−45] To this end, the model system was divided in two parts: (1) the active site region of the enzymatic complex, which was treated at the QM CP level with the DFT-BLYP functional,[46,47] and (2) the remaining protein atoms and water, which were treated at the classical MD level with the AMBER force field. The use of the generally more reliable hybrid B3LYP[48] functional is unfortunately not possible for the present CP-MD simulations because of prohibitively heavy computational cost (about 2 orders of magnitude effort) associated with the use of B3LYP in CPMD.

In detail, the QM part of the system includes the Zn^{2+} ion, the side chains of the coordinated residues (namely, Asp297β, Cys299β, His362β), the cysteine residue of the CaaX sequence, the diphosphate group and the first isoprene subunit of the FPP substrate, the side chains of the surrounding hydrogen donor residues (namely, Lys164α, Arg291β, Lys294β and Tyr300β), and one of the water molecules around the Zn^{2+} coordination shell. In total, 101 atoms were treated at the QM level (Figure 1B). A 25 Å × 20 Å × 20 Å supercell was used for the QM-CP system. The interaction between the valence electrons and the ionic cores are described with Troullier−Martins norm-conserving pseudopotenitals,[49] and a 70 Ry cutoff energy was applied. Simulations are carried out with a fictitious electron mass of 1000 au and a time step of 5 au (0.121 fs). The adiabaticity of the system was checked and assured (see Supporting Information for details). The interactions between the QM and MM regions are treated as in ref 39, and a rigorous treatment of the electrostatic interaction is implemented as in ref 50. The system was coupled with a Nosé−Hoover thermostat at 500 cm^{-1} frequency to achieve constant temperature simulations.[51,52]

The protocol of the QM/MM calculations includes an initial equilibration of the MD starting configuration. First, a short MD simulation was performed where the QM part is kept frozen, while the MM part is free to move for ca. 500 steps. Then, the whole system is allowed to move and gradually heat up to 300K in 1 ps. Finally, 1 ps of free QM/MM CP-MD was performed in order to equilibrate the

system, and provide a configuration to initiate the subsequent constrained QM/MM calculations. The enzymatic mechanism was investigated using a reaction coordinate (RC) defined as the distance between the two reactive atoms, namely the carbon C_1 on the FPP molecule and the S^- anion on the cysteine thiolate (Figure 1B). This RC was able to provide a fair description of the mechanism of nucleophilic substitution. An alternative choice of RC, defined as the difference between the length of the forming bond and that of the breaking bond, was also chosen to describe the reaction mechanism. Unfortunately, the latter RC failed to offer a reasonable picture of the reaction pathway (see Supporting Information for details).

So-called blue-moon ensemble simulations are performed for the model systems.[53] A constraint is applied at different values of the RC, whereas all other degrees of freedom are free to evolve. The catalytic reaction pathways are characterized in terms of (1) free energy profiles calculated using thermodynamic integration;[53] (2) variation of critical bond lengths, averaged over the last 1.5 ps of each constrained CP-MD QM/MM simulation; (3) variation of the electrostatic potential (D-RESP)[54] charges, calculated for all QM atoms during the QM/MM simulations, on the fly, and averaged over the last 1.5 ps of dynamics. Each step is simulated for at least 3 ps, or until the force on the constraint is equilibrated (i.e., the running averages over 1 ps windows varies less than 6%). The free energy profile is obtained by integration of the force profile. The error associated to each point of free energy profiles is calculated by propagating the error on forces at every step, using the propagation of error formula for linear functions. The present estimates of the free energies, based on these short ab initio CP-MD trajectories, should be considered rather approximate. Ideally, longer trajectories and several independent pathways should be investigated for a more accurate estimation of the enzymatic activation free energy, which unfortunately is not possible with currently available computational resources.

Results and Discussion

FPP Peptide Farnesylation. The free energy surface (FES) of the farnesylation reaction is computed using constrained dynamics, as explained in Methodology. Structural changes during the reaction mechanism are described in terms of averaged bond lengths. Twelve windows at different values along the reaction coordinate (RC) are considered in the interval [1.8, 4.0] Å. The RC is the internuclear distance between atoms C_1 and S^- (r_1 in Figure 1B), which represents the bond in formation. The distance between the atom C_1 and the oxygen O_1 of the diphosphate (PPi; r_2), namely the breaking bond, is instead free, together with all the remaining degrees of freedom.

The shape of the resulting FES of the FTase reaction is characterized by two minima, reactant (R) and product (P) states, separated by a single transition state (TS) (Figure 2). The local minimum in the R state is located around $r_1 = 4.0$ Å and $r_2 = 1.52$ Å. This structure is stable during a 2 ps free CP-MD QM/MM trajectory, which is consistent with the conformation produced by preparatory classical MD

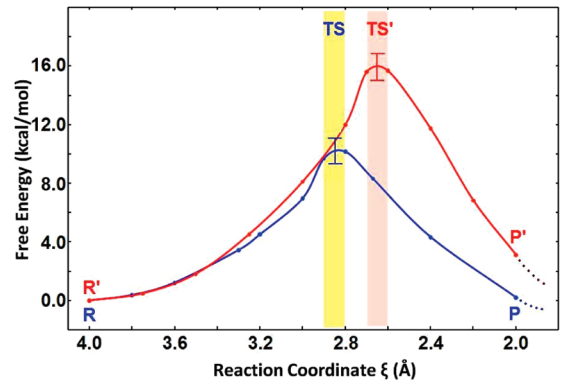


Figure 2. Calculated free energy surface (FES) of the farnesylation reaction. The FES obtained by using the FTase/CH_3-FPP/CaaX complex is shown in blue, while the one obtained by using the FTase/CF_3-FPP/CaaX complex is shown in red. The transition state regions along the two profiles are marked by colored rectangles.

(Figure 3). Simulations of the initial state R show a well-structured H-bond network that stabilizes the FPP binding conformation. Four amino acids are interacting with the substrate, acting as H-bond donors: Lys164α, Arg291β, Lys294β, and Tyr300β. This H-bond network is essential for the stabilization of the PPi and its electrostatic counterbalance.[55] Specifically, β-PPi H-bonds to Lys164α (r_{H1} = 1.51 Å) and to Arg291β (r_{H2} = 2.65 Å), while α-PPi H-bonds to Lys294β (r_{H3} = 1.73 Å) (represented as black lines in Figure 3). Also, the O_1 atom H-bonds to the hydroxyl group on Tyr300β (r_{H4} = 2.26 Å). This conformation of the active site in our simulations shows high similarity with the results previously reported from mutagenesis studies.[24] Interestingly, this H-bond network differs from that found in some X-ray structures and mutagenesis studies[5,23,56] of the *inactive* form, where Lys164α and Arg291β interact with the α-PPi, while Lys294β and Tyr300β interact with β-PPi. The Zn^{2+} metal maintains its starting coordination with residues Asp297β, Cys299β, His362β, and the cysteine thiolate from the CaaX motif. The average bond lengths of this tetracoordination compare well with the crystallographic ones: 2.08 Å (2.08 Å), 2.37 Å (2.21 Å), 2.14 Å (2.17 Å), and 2.35 Å (2.35 Å), respectively, with the crystallographic data shown in the parentheses.[5]

From R, we started to progressively decrease r_1 so as to approach the TS. Within [3.0, 4.0] Å interval no significant structural changes are observed, and the conformation of the farnesyl group, as well as the metal−ligand coordination, remains essentially unchanged (Figure 4). At r_1 = 2.9 Å, the distance between the thiolate and the Zn^{2+} metal ion becomes slightly longer (d_1 = 2.35 Å at r_1 = 4 Å versus 2.39 Å at r_1 = 2.9 Å). Importantly, this event precedes the nucleophilic attack of the thiolate on C_1, while r_2 increases from 1.52 Å (R state) to 1.57 Å. Also, at r_1 = 2.9 Å, resonance of bonds C_1−C_2 (r_3) and C_2−C_3 (r_4) starts to be evidenced by structural changes: the bond C_1−C_2 (r_3) decreases from 1.50 Å to 1.45 Å, while bond C_2−C_3 (r_4) becomes longer, from 1.35 Å to 1.37 Å, and thus suggest the approach of the TS region.

The TS region is located at 2.8 Å < r_1 < 2.9 Å, in which the averaged forces on the constraint are zero (Figure 2 and

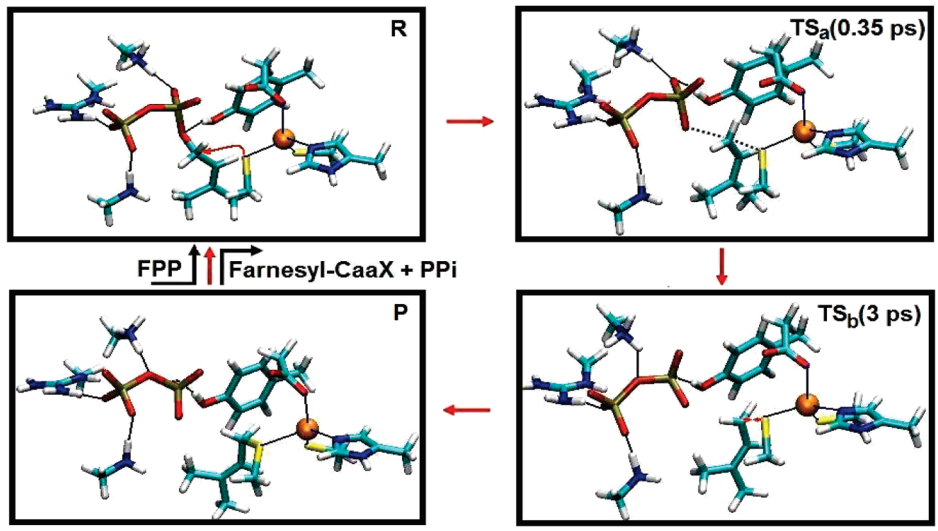


Figure 3. Selected snapshots taken from the reactive pathway of the farnesylation reaction. The metal−ligand coordination is indicated by blue lines. Hydrogen bonds are indicated by black lines. (R) Starting structure of the FTase/CH_3-FPP/CaaX complex. The Zn^{2+}-coordinated thiolate plays a role as a nucleophilic group; (TS_a) selected structure at the transition state ($r_1 = 2.8$ Å) after 0.35 ps of dynamics. Here a modest dissociation of C−O bond (r_2), and a resonance of C_1−C_2 and C_2=C_3 bonds are observed. This points to a fairly dissociative character at the transition state; TS_b) selected structure at the transition state ($r_1 =$ 2.8 Å) after ∼3 ps of dynamics. Now the diphosphate group dissociates completely from the farnesyl group; (P) the final product of the reaction is shown, where the new bond (C−S) is fully formed, indicating completion of the catalytic action.

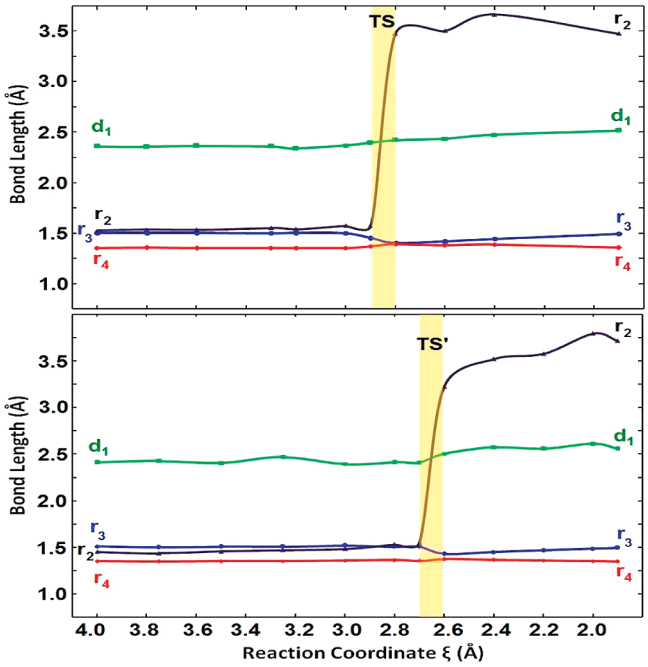


Figure 4. Selected average bond distances (label are indicated in Figure 1B) of the active site in the FTase/FPP/CaaX ternary complex along the investigated reaction pathways. (Upper Panel): FTase/CH_3-FPP/CaaX ternary complex. (Lower Panel): FTase/CF_3-FPP/CaaX ternary complex. The transition states are marked by colored rectangles.

Figure 3). Here, d_1 further elongates to 2.42 Å, indicating that the thiolate is dissociating from the zinc ion and is approaching C_1. At this point, r_2 spontaneously elongates (3.48 Å in TS versus 1.52 Å in R), leading to the definitive bond dissociation of r_2 and the subsequent carbocation formation. Then, the free PPi forms a stable hydrogen bond network with Lys164α, Arg291β, Lys294β, and Tyr300β. Here, we observe a modification of the H-bond network around the PPi group, compared to that in R: Lys294β is now H-bonding to the linking O atom ($r_{H5} = 2.34$ Å) and Tyr300β is H-bonding to the α-PPi ($r_{H4} = 1.74$ Å), while Lys164α and Arg291β still H-bond to β-PPi ($r_{H1} = 1.47$ Å and $r_{H2} = 1.88$ Å). Overall, all H-bonds become shorter when binding to the free PPi group, indicating a stronger electrostatic interaction. Also, r_3 decreases its length further, from 1.45 Å to 1.40 Å, and r_4 becomes slightly longer, from 1.37 Å to 1.39 Å. Although small in magnitude, the decreasing of r_3 together with the increasing of r_4 suggest resonance of the C_1−C_2 and C_2=C_3 bonds in the TS region, which indicates the likely formation of the carbocation on C_1. Nevertheless, we did not observe in the simulations a stable intermediate state in this region, indicating that a stable isolated carbocation is missing. Although we cannot definitely rule out the presence of such a stable intermediate, our findings suggest a reaction with an S_N2-like mechanism having a rather dissociative character, in agreement with the mechanism proposed by Fierke et al.[24,29] Also, our results are further confirmed by a recent study of Lenevich et al., who have proposed a similar TS structure of a nonenzymatic reaction, based on computational calculations and kinetic isotope effect (KIE) studies of yeast FTase.[57] Finally, we point out that solvent water molecules do not seem to play an active role in the catalysis. In fact, there are few water molecules surrounding these two reactive atoms (C_1 and S^-) at the TS region. Specifically, no water molecules are found within a 3 Å sphere centered on these two atoms, and only three within a 5 Å sphere, during catalysis.

After $r_1 = 2.8$ Å, the force on the constraint changes sign, indicating that the system is evolving toward the P state (Figure 3). The constraint is released after 3 ps of CP-MD QM/MM simulation at $r_1 = 2.8$ Å and a stable product structure is formed. Here, $r_1 = 1.9$ Å, while $r_2 = 3.48$ Å.

The final structure agrees well with the crystallographic results;[22] rmsd = 1.8 Å for the entire protein backbone and 0.7 Å for the active site residues. Moreover, the Zn^{2+} ion still maintains its coordination with Asp297β, Cys299β, and His 362β, and distances compare well with the crystallographic data shown in square brackets:[22] 2.03 [2.06] Å, 2.32 [2.27] Å, and 2.11 [2.18] Å, respectively. On the other hand, d_1 is now longer than in R (2.52 [2.66] Å in P[22] versus 2.36 Å in R), indicating that the formation of r_1 weakens the Zn^{2+}−S coordination, likely induced by a weaker charge interaction as discussed below in Charge Evolution during Catalysis.

CF_3-Substituted FPP Peptide Farnesylation. Fluorine substitutions on one methyl group of the alkyl chain of the FPP substrate have been studied to clarify the farnesylation reaction mechanism. The reaction rate drastically decreases when FPP analogues (CF_3-FPP) are used as substrates (from 770-fold[28] up to 3000-fold[29]). This indicates that the trifluoromethyl group hinders the reaction and increases its energy barrier, suggesting a possible electrophilic mechanism. In order to investigate this hypothesis, we generated an FTase/CF_3-FPP/KCVIM complex within the present CP-MD QM/MM protocol. The setup employed for the QM system is identical to the wild-type study. Also, the procedure to define the FES remains the same.

With this analogue system, the FES shows two minima, reactant (R′) and product (P′) states, and one transition state (TS′). The R′ state is stable around $r_1 \sim 4.0$ Å, as for the wild-type system (Figure 2). The catalytic site contains a large hydrophobic cavity that can easily accommodate the CF_3-FPP analogue; the diphosphate group forms a stable hydrogen bond network with Lys164α, Lys294β, and Tyr300β. In details, Lys164α H-bonds to β-PPi ($r_{H1} \sim 1.96$ Å), Lys294β interacts both with α-PPi ($r_{H3} \sim 1.76$ Å) and β PPi ($r_{H6} \sim 1.96$ Å), and Tyr300β H-bonds to O_1 atom ($r_{H4} \sim 2.02$ Å). Here, we could not observe an H-bond between Arg291β and PPi; this distance is larger than 3.5 Å. Overall, the Zn^{2+} tetracoordination is consistent with that in the wild-type system.

The structural conformation at the R′ state is maintained from $r_1 = 4.0$ Å to 2.7 Å. The active site structure, including the farnesyl group and the metal−ligand coordination, does not change within this r_1 interval. The cysteine thiolate coordination to the Zn^{2+} ion is maintained as shown by d_1, which does not change much (∼2.41 Å in R′). Conversely, r_2 changes from 1.50 Å at $r_1 = 4.0$ Å to 1.61 Å at $r_1 = 2.7$ Å, reproducing the same trend as in the wild-type system, but with a larger amplitude. Unlike the resonance event we observed in the wild-type system, where r_3 and r_4 moderately change at $r_1 = 2.9$ A, here r_3 and r_4 do not show significant changes (the differences between r_3 and r_4 is ∼0.01 Å at $r_1 = 4.0$ Å and $r_1 = 2.7$ Å). This indicates that the fluorine substitutions on the methyl group hinder the resonance effect and destabilize the possible formation of a carbocation.

The TS′ region is located at 2.6 Å < r_1 < 2.7 Å, where the averaged forces on the constraint are zero within the statistical error. Compared to the wild-type system, the TS′ region is shifted by 0.2 Å. As for the wild-type, we could not observe a stable intermediate state. In terms of structural changes, d_1 is 2.50 Å, showing that the nucleophile is actually approaching C_1. Concomitantly, the r_2 bond spontaneously breaks (3.41 Å, Figure 4). This clearly indicates the dissociation of the C−O bond. Also, r_3 decreases to 1.43 Å and r_4 elongates to 1.37 Å. This evidences a resonance structure that is similar to the wild-type case. At the same time, the dissociated PPi group forms H-bond interactions with Lys164α, Lys294β, and Tyr300β. Now the $r_{H1} = 1.77$ Å, $r_{H3} = 1.87$ Å, $r_{H6} = 1.64$ Å, and $r_{H4} = 1.63$ Å. Only r_{H3} is slightly longer than in R′, whereas the rest of the values are shorter than in R′.

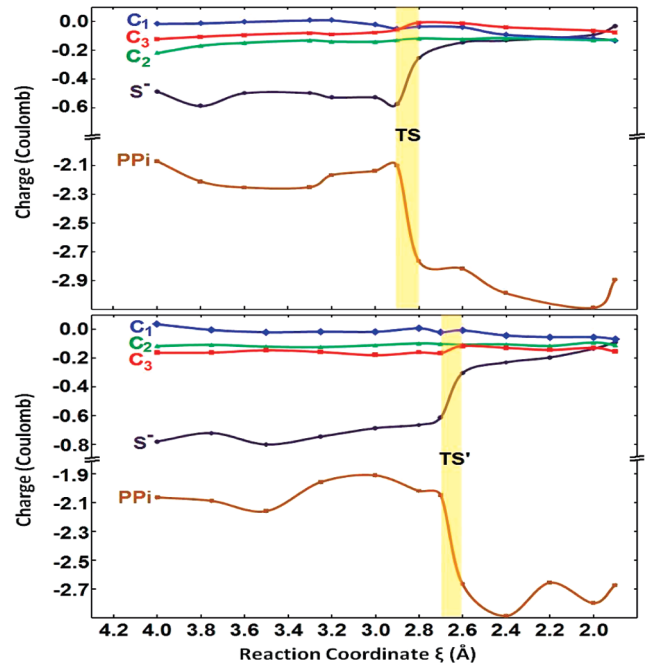


Figure 5. Profile of the ESP charge distribution on the active atoms and groups (labels of atoms/groups are indicated in Scheme 1) in the active site along the investigated reaction pathways. (Upper panel): FTase/CH_3-FPP/CaaX ternary complex; (Lower Panel): FTase/CF_3-FPP/CaaX ternary complex. The transition states are marked by colored rectangles.

At $r_1 = 2.6$ Å, the force on the constraint changes sign, indicating that the system is falling into the P′ well. The constraint is then released, and the system freely falls into the product well. The average bond length of C_1−S is 1.9 Å. In the P′ region, the dissociated phosphate group continues to form a hydrogen bond network with the surrounding residues Lys164α, Lys294β, and Tyr300β. Arg291β does not play a role of H-bond donor, as previously observed. The conformation of the CF_3-FPP-farnesylated KCVIM peptide and the structure of the metal ion coordination match the crystallographic structure of the nonsubstituted product.[22]

Charge Evolution during Catalysis. Both reaction pathways can be monitored through charge variations of relevant molecular moieties, as reported in Figure 5. The D-RESP atomic charge[54] was assigned to each QM atom, based on the electrostatic potential in the CP-MD QM/MM simulation computed on the fly.

In the wild-type system, at the initial state R, charge transfer effects occur at the metal−ligand coordination sphere: Zn^{2+} metal ion ($q = +0.16$ in units of electron charge) accepts electron density from Asp297β, Cys299β,

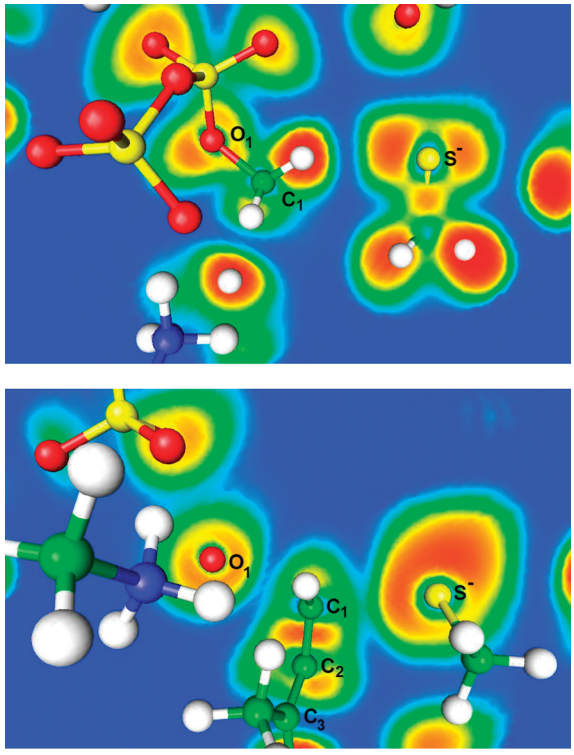


Figure 6. The calculated electron localization function (ELF) plot of the active site in the reagent state (upper) and the transition state (lower). The contour profiles are drawn through the plane of the O_1 and C_1 atoms the and S^- anion.

His 362β, and cysteine thiolate ($q = -0.53$ on the S^- anion). Also, the C_1 atom is virtually neutral ($q = -0.01$) and both C_2 and C_3 atoms are negative ($q = -0.15$ and $q = -0.09$, respectively). A high electron density located on the PPi group reflects its negative charge ($q = -2.20$) and is stabilized by the H-bond interactions that involve the PPi group. At $r_1 = 2.9$ Å, the cysteine thiolate gains electron density ($q = -0.57$) due to its partial dissociation from the metal ion. This shows the nucleophilic character of the thiolate. Here, PPi does not dissociate yet and the local electron density remains the same. At the TS region, PPi spontaneously dissociates and gains electron density ($q = -2.76$). Meanwhile, the decrease of electron density on C_2═C_3 bond is concomitant with the neutralization of the charge of C_1 and C_3 atoms: this evidences a resonance structure formed by C_1, C_2, and C_3 atoms in TS, which is also supported by structural changes discussed earlier (i.e., change of r_3 and r_4 lengths). In particular, the resonance structure shows a partial positive charge transferred on these three atoms ($q = -0.24$ at R to -0.15 at TS). This finding supports a dissociative character with partial positive charge formed in the TS region. Moreover, a charge transfer occurs between the thiolate ($q = -0.25$) and the C_1 atom ($q = -0.03$), showing that the forming positive charge is shared between C_1 and S^- and therefore indicating that the thiolate is attacking the C_1 atom. The calculation of the electron localization function[58] (ELF) also shows that a covalent bond between the sulfur anion and the carbon atom C_1 (Figure 6) starts to form at the TS region. This explains the decrease of the negative charge on the nucleophile. The observation of the resonance of bond r_3, r_4, together with the calculation of the atomic charge and ELF calculation, reveals that the reaction mechanism is either a pure associated or a dissociated mechanism. The structural and electronic property of the transition state in our QM/MM simulation suggests that a weak covalent bond forms between C_1 and S^-. At the same time, the resonance of the group C_1, C_2, and C_3 stabilize the forming partial positive charge, which is due to the dissociation of the diphosphate group. Interestingly, this result is similar to several experimental findings that support a mechanism like the one described above.[20,24,29,33] In P, the thiolate loses electron density while the r_1 progressively decreases, and the atomic charge on S^- anion becomes $q = -0.03$ at the final point. This indicates that the thiolate now is bound to C_1 and coordinated to the Zn^{2+} at the same time. Also, C_1 gains electron density from the thiolate and the charge becomes $q = -0.13$. Finally, C_2 and C_3 gain back electron density, although to a lesser extent compared to that in R ($q = -0.13$ and -0.07, respectively).

The charge distribution along the reaction in the fluorinated system is virtually identical to that in the wild-type. In R′, the electron density on C_1 and C_2 ($q = 0.00$ and -0.11, respectively) is similar to that in the wild-type system. Also, C_3 gains electron density from the nearby triflouromethyl group ($q = -0.16$). The charge distribution does not change when r_1 decreases from 4.0 Å to 2.7 Å. It is worth noting that upon reaching TS′, the electron density on C_1, C_2, and C_3 does not change significantly ($q = -0.27$ to $q = -0.23$). This finding might be explained by the electrostatic effect of the CF_3-FPP. Finally, in the P′ state, the farnesylated cysteine thiolate still binds to the zinc ion and the charge on the sulfur anion is $q = -0.08$, similar to the wild-type system. Therefore, the main difference between TS and TS′ states is the charge distribution, where the C_1, C_2, and C_3 resonance group shows less positive charge in the fluorinated system. This possibly destabilizes the transition state and thereby induces the observed increase of the energy barrier.

Energetics of the Enzymatic Reactions. The free energy profile of different systems (FPP-KCVIM peptide and CF_3-substituted FPP-KCVIM peptide) are computed by thermodynamic integration of the constrained forces along the RC, as described the Methodology and plotted in Figure 2. The free energy barrier in the wild-type system is 10.8 ± 1.0 kcal/mol and 17.2 ± 1.0 kcal/mol in the CF_3-FPP system. Based on transition state theory, the reaction rate in the wild-type system is approximately 4 orders of magnitude faster than that in the fluorinated one. Given the uncertainty related to the calculated free energies and the fact that different CaaX motifs (CVLS[29] and CVIA[28]) were adopted in the experimental measurement, this result agrees fairly well with experimental data,[28,29] which report that the reaction rate for the CF_3-FPP system is slower than the wild-type by 3 to 4 orders of magnitude.

The experimental reaction rate is 0.017 s^{-1} in an FPP-GCVLS peptide system[29] and is 0.0026 s^{-1} in FPP-TKCVIF peptide system,[59] values that correspond to a barrier of 20.0 and 21.1 kcal/mol, respectively. These experimental values actually refer to the complete catalysis, the physical step (i.e., the conformational change of FPP) followed by the chemical step (i.e., the farnesylation reaction), while this study is

focused only on the latter. Thus, a direct quantitative comparison of experimental and theoretical enzymatic activity is difficult. Nevertheless, it is worth mentioning that a recent DFT study, which was carried out with a different procedure compared to the one applied herein, has reported that the cost of bringing the two reactive groups close to each other to overcome the physical step barrier is 10 kcal/mol.[26] If one assumes that the two steps (physical and chemical) are additive, then there seems to be good agreement between the computed and experimental barriers. The possibility that the barriers of the physical and chemical steps are additive is suggested by the lack of the "intermediate" state in the active reagents, as evidenced by several experimental findings such as spectroscopy studies and crystallographic data.[2,22] This might indicate that, between the physical and chemical steps, the potential minimum is either missing (fully additive barriers) or relatively small. At this point, however, this rough comparison is somewhat speculative. Time-consuming accurate calculations on the physical step with the present methodology are currently in progress that will allow a correct comparison of the calculated enzymatic barrier with the experimental values.

The shape of the FES of the chemical step shows that the enzymatic reaction essentially follows an associative mechanism (i.e., no intermediate), where the TS is located between $r_1 = 2.8$ Å and 2.9 Å. We observe a significant resonance structure involving C_1, C_2, and C_3 atoms, together with a complete dissociation of the phosphate group at the TS region, which however does not lead to the formation of a stable intermediate. This can be explained by examining the structure, charge distribution, and the ELF plot at the TS region. As already mentioned in the previous section, a charge transfer occurred between the two reactive atoms, C_1 and S^-. This indicates that the cysteine thiolate is interacting covalently with C_1 during the nucleophilic attack, in the TS. Therefore, though we observe dissociative characteristics along this reaction pathway, the overall chemical step should not be classified as a pure S_N1 mechanism.

The fluorinated system generally follows the same pattern, except from the fact that the TS′ is at r_1 interval [2.6, 2.7] Å, where less resonance effects are observed. The free energy cost to bring r_1 from 4.0 Å to 2.8 Å is the same as that in the wild-type system, showing that the substitution of the methyl side chain on FPP does not introduce additional steric effects. Instead, the TS′ state is destabilized due to the electron-withdrawing effect of the trifluoromethyl group and therefore results in a higher free energy barrier and a late TS event. This phenomenon is consistent with the observed structural information and charge transfer effect.

Conclusion

FTase has become a popular research subject since the discovery of the relation between its oncogenesis peptide substrates and the development of human cancers. To date, many potential FTase inhibitors have been extensively developed and showed encouraging preclinical results.[60] Some, such as tipifarnib and lonafarnib, were tested in human clinical trials.[61−64] Moreover, recent studies point out that FTase inhibitors, originally considered only as anticancer agents, show promising effects in treating malaria.[65−68] As a result, FTase is currently a target in drug discovery and development.

Numerous crystallographic and kinetic studies have been performed for FTase, and different catalytic reaction pathways have been proposed. In this article, we have presented an investigation of the mechanism and energetics of the enzymatic reaction catalyzed by the FTase. In addition, the inhibitory effect of fluorinated substrate has been investigated, using a substrate analogue constituted by a CF_3-FPP.

The present simulations indicate the enzymatic reaction occurs via the so-called "associative mechanism with dissociative character", in agreement with the proposed models based on experimental data.[29,33,57] We observe a resonance structure in the TS region, which is concomitant with the formation of a metastable carbocation. Charge transfer effects along the reaction pathway confirm the resonance structure in the TS region. Nevertheless, no stable intermediate is found during the catalysis, suggesting a single-step mechanism. Inversion of configuration of the carbocation is also observed, in agreement with an S_N2-like mechanism. The dissociative character is explained by the fairly long length of the bond in breaking (3.5 Å) in the transition state, while the bond formation has a value of 2.8 Å.

The free energy for the chemical step of the catalysis is 10.8 ± 1.0 kcal/mol. Interestingly, fluorine substitution of FPP (CF_3-FPP) increases the energy barrier to 17.2 ± 1.0 kcal/mol, in agreement with the experimental measurements[28,29] despite the fact that the reaction mechanism remains similar to that found with the FPP substrate. Charge transfer effects inducing destabilization of the carbocation formation in the CF_3-FPP reaction mechanism is the major reason for the increase of the barrier: the electron-withdrawing fluorine atoms hindered the reaction, leading to a higher energy for the transition state.

In summary, through the use of CP-MD QM/MM computations, we have proposed a picture of the FTase reaction mechanism and shed light on the inhibitory effect of fluorine substituents of the FPP substrate. The present description of the transition state conformations along the catalytic pathways might be helpful in the design of selective FTase inhibitors.

Acknowledgment. We thank the National Institutes of Health (NIH) for financial support and the Pittsburgh Supercomputer Center (PSC) for providing computational resources.

Supporting Information Available: Further details are given concerning the classical MD simulations, the QM/MM setup, the RESP charges used for the catalytic residues in the active site, the rmsd values of the enzyme along the trajectory, and snapshots illustrating the alternative pathways investigated. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Zhang, F. L.; Casey, P. J. *Annu. Rev. Biochem.* **1996**, *65*, 241–269.

(2) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Biol. Inorg. Chem.* **2005**, *10*, 3–10.

(3) Lane, K. T.; Beese, L. S. *J. Lipid Res.* **2006**, *47*, 681–699.

(4) Kato, K.; Cox, A. D.; Hisaka, M. M.; Graham, S. M.; Buss, J. E.; Der, C. J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 6403–6407.

(5) Long, S. B.; Hancock, P. J.; Kral, A. M.; Hellinga, H. W.; Beese, L. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12948–12953.

(6) Reiss, Y.; Stradley, S. J.; Gierasch, L. M.; Brown, M. S.; Goldstein, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 732–736.

(7) Hartman, H. L.; Hicks, K. A.; Fierke, C. A. *Biochemistry* **2005**, *44*, 15314–15324.

(8) Vojtek, A. B.; Der, C. J. *J. Biol. Chem.* **1998**, *273*, 19925–19928.

(9) Reuter, C. W. M.; Morgan, M. A.; Bergmann, L. *Blood* **2000**, *96*, 1655–1669.

(10) Mazieres, J.; Pradines, A.; Favre, G. *Cancer Lett.* **2004**, *206*, 159–167.

(11) Brunner, T. B.; Hahn, S. M.; Gupta, A. K.; Muschel, R. J.; McKenna, W. G.; Bernhard, E. *J. Cancer Res.* **2003**, *63*, 5656–5668.

(12) Barbacid, M. *Annu. Rev. Biochem.* **1987**, *56*, 779–827.

(13) Takai, Y.; Sasaki, T.; Matozaki, T. *Physiol. Rev.* **2001**, *81*, 153–208.

(14) Reiss, Y.; Goldstein, J. L.; Seabra, M. C.; Casey, P. J.; Brown, M. S. *Cell* **1990**, *62*, 81–88.

(15) Furfine, E. S.; Leban, J. J.; Landavazo, A.; Moomaw, J. F.; Casey, P. J. *Biochemistry* **1995**, *34*, 6857–6862.

(16) Pompliano, D. L.; Schaber, M. D.; Mosser, S. D.; Omer, C. A.; Shafer, J. A.; Gibbs, J. B. *Biochemistry* **1993**, *32*, 8341–8347.

(17) Long, S. B.; Casey, P. J.; Beese, L. S. *Structure* **2000**, *8*, 209–222.

(18) Reiss, Y.; Brown, M. S.; Goldstein, J. L. *J. Biol. Chem.* **1992**, *267*, 6403–6408.

(19) Hightower, K. E.; Huang, C. C.; Casey, P. J.; Fierke, C. A. *Biochemistry* **1998**, *37*, 15555–15562.

(20) Saderholm, M. J.; Hightower, K. E.; Fierke, C. A. *Biochemistry* **2000**, *39*, 12398–12405.

(21) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Mol. Struct. (THEOCHEM)* **2005**, *729*, 125–129.

(22) Long, S. B.; Casey, P. J.; Beese, L. S. *Nature (London)* **2002**, *419*, 645–650.

(23) Reid, T. S.; Long, S. B.; Beese, L. S. *Biochemistry* **2004**, *43*, 6877–6884.

(24) Pickett, J. S.; Bowers, K. E.; Hartman, H. L.; Fu, H. W.; Embry, A. C.; Casey, P. J.; Fierke, C. A. *Biochemistry* **2003**, *42*, 9741–9748.

(25) Harris, C. M.; Derdowski, A. M.; Poulter, C. D. *Biochemistry* **2002**, *41*, 10554–10562.

(26) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 205–218.

(27) Cassidy, P. B.; Poulter, C. D. *J. Am. Chem. Soc.* **1996**, *118*, 8761–8762.

(28) Dolence, J. M.; Poulter, C. D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 5008–5011.

(29) Huang, C. C.; Hightower, K. E.; Fierke, C. A. *Biochemistry* **2000**, *39*, 2593–2602.

(30) Mu, Y. Q.; Omer, C. A.; Gibbs, R. A. *J. Am. Chem. Soc.* **1996**, *118*, 1817–1823.

(31) Weller, V. A.; Distefano, M. D. *J. Am. Chem. Soc.* **1998**, *120*, 7975–7976.

(32) Pearson, R. G. *J. Am. Chem. Soc.* **1963**, *85*, 3533–3539.

(33) Wu, Z.; Demma, M.; Strickland, C. L.; Radisky, E. S.; Poulter, C. D.; Le, H. V.; Windsor, W. T. *Biochemistry* **1999**, *38*, 11239–11249.

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

(35) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(36) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(37) Cui, G.; Wang, B.; Merz, K. M. *Biochemistry* **2005**, *44*, 16513–16523.

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

(39) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.

(40) Dal Peraro, M.; Ruggerone, P.; Raugei, S.; Gervasi, F. L.; Carloni, P. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149–156.

(41) Dal Peraro, M.; Llarrull, L. I.; Rothlisberger, U.; Vila, A. J.; Carloni, P. *J. Am. Chem. Soc.* **2004**, *126*, 12661–12668.

(42) De Vivo, M.; Ensing, B.; Klein, M. L. *J. Am. Chem. Soc.* **2005**, *127*, 11226–11227.

(43) De Vivo, M.; Ensing, B.; Dal Peraro, M.; Gomez, G. A.; Christianson, D. W.; Klein, M. L. *J. Am. Chem. Soc.* **2007**, *129*, 387–394.

(44) De Vivo, M.; Cavalli, A.; Carloni, P.; Recanatini, M. *Chem.—Eur. J.* **2007**, *13*, 8437–8444.

(45) De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 10955–10962.

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

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

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

(49) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 8861–8869.

(50) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.

(51) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

(52) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(53) Ciccotti, G.; Ferrario, M.; Hynes, J. T.; Kapral, R. *Chem. Phys.* **1989**, *129*, 241–251.

(54) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.

(55) Bowers, K. E.; Fierke, C. A. *Biochemistry* **2004**, *43*, 5256–5265.

(56) Strickland, C. L.; Windsor, W. T.; Syto, R.; Wang, L.; Bond, R.; Wu, Z.; Schwartz, J.; Le, H. V.; Beese, L. S.; Weber, P. C. *Biochemistry* **1998**, *37*, 16601–16611.

(57) Lenevich, S.; Xu, J.; Hosokawa, A.; Cramer, C. J.; Distefano, M. D. *J. Am. Chem. Soc.* **2007**, *129*, 5796–5797.

(58) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5379–5403.

(59) Pais, J. E.; Bowers, K. E.; Fierke, C. A. *J. Am. Chem. Soc.* **2006**, *128*, 15086–15087.

(60) Bell, I. M. *J. Med. Chem.* **2004**, *47*, 1869–1878.

(61) Doll, R. J.; Kirschmeier, P.; Bishop, W. R. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 478–486.

(62) Kurzrock, R.; Kantarjian, H. M.; Blascovich, M. A.; Bucher, C.; Verstovsek, S.; Wright, J. J.; Pilat, S. R.; Cortes, J. E.; Estey, E. H.; Giles, F. J.; Beran, M.; Sebti, S. M. *Clin. Cancer Res.* **2008**, *14*, 509–514.

(63) Adjei, A. A.; Erlichman, C.; Davis, J. N.; Cutler, D. L.; Sloan, J. A.; Marks, R. S.; Hanson, L. J.; Svingen, P. A.; Atherton, P.; Bishop, W. R.; Kirschmeier, P.; Kaufmann, S. H. *Cancer Res.* **2000**, *60*, 1871–1877.

(64) Ready, N. E.; Lipton, A.; Zhu, Y.; Statkevich, P.; Frank, E.; Curtis, D.; Bukowski, R. M. *Clin. Cancer Res.* **2007**, *13*, 576–583.

(65) Gelb, M. H.; Hol, W. G. J. *Science* **2002**, *297*, 343–344.

(66) Bulbule, V. J.; Rivas, K.; Verlinde, C. L.; Van Voorhis, W. C.; Gelb, M. H. *J. Med. Chem.* **2008**, *51*, 384–387.

(67) Nallan, L.; Bauer, K. D.; Bendale, P.; Rivas, K.; Yokoyama, K.; Hornéy, C. P.; Pendyala, P. R.; Floyd, D.; Lombardo, L. J.; Williams, D. K.; Hamilton, A.; Sebti, S.; Windsor, W. T.; Weber, P. C.; Buckner, F. S.; Chakrabarti, D.; Gelb, M. H.; Van Voorhis, W. C. *J. Med. Chem.* **2005**, *48*, 3704–3713.

(68) Olepu, S.; Suryadevara, P. K.; Rivas, K.; Yokoyama, K.; Verlinde, C.; Chakrabarti, D.; Van Voorhis, W. C.; Gelb, M. H. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 494–497.

CT8004722

JCTC Journal of Chemical Theory and Computation

Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

Departamento de Química Física y Analítica, Universidad de Oviedo, 33006 Oviedo (Asturias), Spain

Received November 17, 2008

Abstract: Herein, we first review different methodologies that have been proposed for computing the quantum mechanical (QM) energy and other molecular properties of large systems through a linear combination of subsystem (fragment) energies, which can be computed using conventional QM packages. Particularly, we emphasize the similarities among the different methods that can be considered as variants of the multibody expansion technique. Nevertheless, on the basis of thermochemical arguments, we propose yet another variant of the fragment energy methods, which could be useful for, and readily applicable to, biomolecules using either QM or hybrid quantum mechanical/molecular mechanics methods. The proposed computational scheme is applied to investigate the stability of a triple-helical collagen model peptide. To better address the actual applicability of the fragment QM method and to properly compare with experimental data, we compute average energies by carrying out single-point fragment QM calculations on structures generated by a classical molecular dynamics simulation. The QM calculations are done using a density functional level of theory combined with an implicit solvent model. Other free-energy terms such as attractive dispersion interactions or thermal contributions are included using molecular mechanics. The importance of correcting both the intermolecular and intramolecular basis set superposition error (BSSE) in the QM calculations is also discussed in detail. On the basis of the favorable comparison of our fragment-based energies with experimental data and former theoretical results, we conclude that the fragment QM energy strategy could be an interesting addition to the multimethod toolbox for biomolecular simulations in order to investigate those situations (e.g., interactions with metal clusters) that are beyond the range of applicability of common molecular mechanics methods.

Introduction

The idea of representing the total energy of a large molecule as a combination of fragment energies has been considered for decades. To better appreciate their similarities and differences, we will first review several computational approaches for combining fragment energies that have been developed during recent years. We note, however, that other linear-scaling methodologies[1,2] aimed at construction of the full density matrix of a large system from the fragment density submatrices are beyond the scope of this paper. Thus, we will discuss first the methods based on the multibody expansion approach and other closely related methods that include implicitly high-order many-body effects into fragment energies using various approximations. We will also comment on the so-called kernel energy method that turns out to be essentially a multibody expansion method. Subsequently, we will review other methods that approximate the quantum mechanical energy of large systems by combining fragment energies on the basis of intuitive and/or thermochemical argumentations. Although we will see that these thermochemically based protocols can be considered as truncated forms of the more general multibody expansion method, they are conceptually simpler and can be readily

* Corresponding author phone: +34-985103689; fax: +34-985103125; e-mail: dimas@uniovi.es.

10.1021/ct8005002 CCC: $40.75
Published on Web 04/29/2009

applicable using many computational tools at a moderate computational cost. In fact, we will formulate yet another variant of the thermochemical fragment energy methods that could be particularly useful to compute the energies of large biomolecular systems. Finally, as a real case application of the proposed method, we will combine fragment-based quantum chemical energies with molecular mechanics and standard quantum chemical calculations in order to compute the relative free energy of the triple-helical form of a collagen model peptide with respect to its monomer state.

Multibody Expansion Method. The so-called cluster expansion method[3] has been developed in the framework of solid-state chemistry in order to represent the total energy of an atomic crystal as a linear combination of the characteristic energies of clusters of atoms over a fixed lattice. The coefficients in the cluster expansion are computed using quantum mechanical energy calculations of a few prototype structures. However, the so-constructed functions are not transferable, i.e., they cannot be used for each conceivable configuration of the system. Subsequently, the multibody expansion (MBE) method, also called *N*-body potentials, or otherwise, cluster potentials, has been developed as a more refined version of the cluster expansion technique.[4] The MBE method evaluates the total energy as a summation of energies corresponding to isolated atomic clusters extracted from the global structure so that they include systematically two-, three-, and *N*-body effects. More recently, it has been demonstrated that the MBE approach can be generalized for an *arbitrary* system, whose energy can be uniquely evaluated using series of structure-independent, perfectly transferable, many-body potentials.[5] In this general MBE formalism, the total energy of an *M*-particle system (composed of atoms, molecules, or molecular fragments linked covalently) can be expressed as $E_M(A_1, A_2, ..., A_M)$, where $A_i = \{\mathbf{R}_i, \sigma_i\}$ has the information about the coordinates ($\mathbf{R}_i$) and the type (σ_i) of the i particle. Since the ordering of the M particles is arbitrary, the functional form of E_M must be such that E_M is invariant to any permutation $A_i \leftrightarrow A_j$.

Representing the total energy by an expansion of a series of *N*-order (or *N*-body or *N*-fragment) energy contributions $E^{(N)}$, we have

$$E_M(A_1, A_2, ..., A_M) = \sum_{N=1}^{M} E^{(N)}(A_1, A_2, ..., A_M) \quad (1)$$

where, in turn, the $E^{(N)}$ terms can be computed from a multiple summation of *N*-order interaction potentials

$$E^{(N)} = \sum_{m_1<...<m_N}^{M} V^{(N)}(A_{m_1}, A_{m_2}, ..., A_{m_N}) \quad (2)$$

where the sum $\sum_{m_1<...<m_N}^{M} V^{(N)}$ runs over all possible combinations $\{m_1, ..., m_N\} \in \{1, ..., M\}$.

Note that eqs 1 and 2 express the total energy E in terms of *N*-order potentials. In practice, however, one needs to compute the $V^{(N)}$ potentials from energy calculations performed on different subsystems. The general relationship between $V^{(N)}$ and subsystem energies can be obtained through a Möbius inversion as defined in number theory.[5] The general result is

$$V^{(N)}(A_1, A_2, ..., A_N) = \sum_{L=1}^{N} (-1)^{N-L} \sum_{m_1<...<m_L}^{N} E(A_{m_1}, A_{m_2}, ..., A_{m_L}) \quad (3)$$

In the above equation, $E(A_{m_1}, A_{m_2}, ..., A_{m_L})$ stands for the energy of a cluster composed by L fragments labeled by the $(m_1, m_2, ..., m_L)$ indices. In fact, eq 3 constitutes a unique definition of the *N*-order interaction potential $V^{(N)}$, which is structure independent because this equation does not carry any information about the environment of the subsystems.[5] The actual significance of eq 3 can be more easily grasped by deriving the first terms of the *N*-order expansion leading to the total energy. Thus, the sum of the first-order potentials is just the sum of the energies of the isolated fragments

$$E^{(1)} = \sum_{m_1=1}^{M} V^{(1)}(A_{m_1}) = \sum_{m_1=1}^{M} E(A_{m_1}) \quad (4)$$

For the second-order contribution, which can be interpreted as the *excess* energy due to pair interactions, we obtain

$$E^{(2)} = \sum_{m_1<m_2}^{M} V^{(2)}(A_{m_1}, A_{m_2}) = \sum_{m_1<m_2}^{M} [E(A_{m_1}, A_{m_2}) - E(A_{m_1}) - E(A_{m_2})] \quad (5)$$

and, of course, $E_M \approx E^{(1)} + E^{(2)}$ defines the well-known pairwise additive approximation to the total energy. Analogously, the three-body $E^{(3)}$ contribution, which collects the $V^{(3)}$ potentials, is the additional energy due to three-body effects, and that cannot be assessed from a two-body representation

$$V^{(3)} = \sum_{m_1<m_2<m_3}^{M} [E(A_{m_1}, A_{m_2}, A_{m_3}) - E(A_{m_1}) - E(A_{m_2}) - E(A_{m_3}) - V^{(2)}(A_{m_1}, A_{m_2}) - V^{(2)}(A_{m_1}, A_{m_3}) - V^{(2)}(A_{m_2}, A_{m_3})] \quad (6)$$

Finally, it may be interesting to note that the MBE equations can be rewritten in terms of the so-called mutual information functions (MIFs),[6] which have been used to compute the configurational entropy of flexible molecules. Thus, the MIF expansion approaches the full-dimensional configurational probability distribution by including systematically *N*-order correlations among the internal degrees of freedom; likewise, the successive $V^{(N)}$ potentials include the *N*-order effects on the total energy. Similarly, the energy of a system composed of M arbitrary fragments can be expanded using the MIFs in the following form

$$E_M(A_1, A_2, ..., A_M) = \sum_{i=1}^{M} E(A_i) - \sum_{m_1<m_2}^{M} I_2(A_{m_1}, A_{m_2}) + ... + (-1)^{N-1} \sum_{m_1<...<m_N}^{M} I_N(A_{m_1}, ..., A_{m_N}) \quad (7)$$

where the mutual information function $I_N(A_{m_1}, ..., A_{m_N})$ combines the energies of all the clusters formed by N fragments

$$I_N(A_{m_1}, ..., A_{m_N}) = \sum_{L=1}^{N} (-1)^{L+1} \sum_{m_1<...<m_L}^{N} E(A_{m_1}, ..., A_{m_L}) \quad (8)$$

Note that the mathematical form of the MBE and MIF expressions are identical due to the fact that $(-1)^{N-L} \equiv (-1)^{N+L}$.

Kernel Energy Method is an MBE Method. At this point, it is convenient to simplify the notation used in the MBE equations by replacing $E(A_{m_1}, A_{m_2}, ..., A_{m_L})$ (the energy of the subsystem with L fragments) with $E_{ijk...}$ (the energy of the subsystem composed of the *i, j, k, ...* particles or fragments). In this way, the pairwise additive approximation for a system composed of a total of M fragments can be written as

$$E_M = \sum_{i=1}^{M} E_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} (E_{ij} - E_i - E_j) \quad (9)$$

In recent years, the so-called kernel energy method (KEM) has been utilized to compute the quantum mechanical (QM) energy of large biomolecules[7−11] by representing a full molecule by smaller *kernels* of atoms (i.e., fragments A_i). The majority of the KEM applications that have been reported to date compute the total energy "by summation over the energy contributions of all *double* kernels reduced by those of any *single* kernels, which have been overcounted in the sum over double kernels",[8] that is, by means of the following expression

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M - 2) \sum_{i=1}^{M} E_i \quad (10)$$

However, it can be easily demonstrated (see Supporting Information) that the original KEM energy formula is equivalent to the MBE pairwise additive approximation.

Several KEM applications on biomolecules have been reported in which the dangling bonds of the molecular fragments are saturated with hydrogen atoms before carrying out the corresponding fragment energy calculations. However, the presence of the H-link atoms introduces an error in the computation of the total energy given that the validity of the MBE equations requires that only the actual fragments are considered in the calculations. Nevertheless, if the fragments are large enough and the total number of fragments is relatively low, the associated error can be reasonably small. Of course, the H-link error can be further reduced by including higher order MBE terms given that these terms progressively account for the environment of each fragment by considering larger and larger clusters of fragments. This has been done in a recent article in which the KEM equation is expanded up to fourth order[11] through a cumbersome derivation that follows an MBE recipe employed in a former study of water clusters.[12]

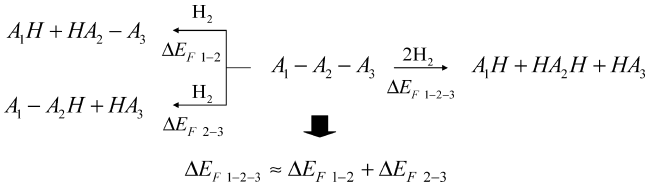
Electrostatically Embedded MBE Methods. In principle, the pairwise additive approximation defined by eq 9 is not enough to accurately compute the total energy of complex systems. Unfortunately, the calculation of higher order MBE terms is extremely expensive in terms of computer time. In order to overcome the limitations of second-order methodologies at a reasonable computational cost, some authors proposed to compute the energies of the individual fragments (E_i) and fragment pairs (E_{ij}) taking into account the electrostatic field of the rest of the system.[13−18] For example, in the fragment molecular orbital (FMO) method, the energies of the different fragments are computed by iteratively solving *effective* fragment Hamiltonians that include the electrostatic effects from the electrons in the surrounding $(M - 1)$ fragments as well as from all nuclei in the total molecule.[14,19] The resulting FMO energies are then combined using MBE equations of order 2 or 3 to derive the total energy. A similar alternative for noncovalently connected fragments is the electrostatically embedded many-body expansion (EE-MBE).[16−18] The energy of each cluster is calculated in the presence of the electric field due to the fixed partial atomic charges of the surrounding fragments. A significant improvement in the electrostatically embedded second- and third-order energies for a series of water clusters is found when compared with the results of standard MBE calculations.[16]

Molecular Tailoring Approach. The so-called molecular tailoring approach (MTA)[20] divides the total system into *overlapping* fragments and subsequently estimates the total energy by summing the fragment contributions and then subtracting the energies of fragment *intersections*. This means that interactions between nonoverlapping fragments are neglected in the MTA method and that each fragment intersection formally accounts for N-body effects to the total energy, with N being the number of overlapping fragments at the particular intersection. This strategy is somehow equivalent to employing localized multibody expansions, and therefore, the MTA approach can be considered as a *flexible* MBE method. The MTA method can also compute one-electron properties of the full system by combining the fragment density matrices into a single density matrix for the whole system.[21]

Molecular Fractionation with Conjugate Caps. The so-called molecular fractionation with conjugate caps (MFCC) scheme also estimates the total energy of large systems from calculations performed on fragments. The MFCC method was originally designed to compute the QM interaction energy between a protein and a small ligand,[22] but this method has been expanded to predict the total energy of protein molecules.[23] In this approach, the protein is divided into fragments $A_i = (-C_\alpha HR_i-CO-N_{i+1}H-)$, with R_i being the side chain of the i amino acid residue and N_{i+1} is the backbone N atom of the $(i + 1)$ amino acid. Instead of H-link atoms, two "conjugate caps", NH_2- and $-C_\alpha H_2R_{i+1}$, are placed at the corresponding $C_{\alpha,i}/N_{i+1}$ atoms to saturate the exposed valence sites of each fragment A_i. The total energy of an M-residue protein molecule is first approximated by summing the energies of the (capped) fragments and then subtracting the energies of the $NH_2-C_\alpha H_2R_{i+1}$ conjugate caps. This first-order approximation is then corrected ad hoc by adding a second-order term ($\delta E^{(2)}$) that accounts for the pairwise interaction energy between non-neighboring fragments. The final MFCC expression is

$$E_M = [E(A_1 - C_\alpha H_2R_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - C_\alpha H_2R_{i+1}) +$$

Scheme 1



$$E(\mathrm{NH_2} - A_M)] - [\sum_{i=1}^{M-1} E(\mathrm{NH_2} - \mathrm{C_\alpha H_2 R}_{i+1})] + \delta E^{(2)} \quad (11)$$

To compute the $\delta E^{(2)}$ contribution, the fragments are capped with H-link atoms as in the KEM scheme. Alternatively, another variant of the MFCC method has been proposed that uses only fragment energies, which are computed in the presence of the electrostatic field created by point charges representing the non-neighboring residues.[24]

Systematic Molecular Fragmentation. As we will see later, the MFCC expression[11] can be justified by means of simple thermochemical arguments on the basis of formal fragmentation processes of the protein system. In fact, the thermochemical approach for computing the fragment-based energy of large molecules has already been explored systematically by Collins et al.[25] The basic reasoning behind the generalization proposed by Collins et al. is summarized in Scheme 1, which shows a generic molecular system composed of three fragments ($A_1-A_2-A_3$) that can be formally broken through three different fragmentation processes.

The key approximation in the protocol of Collins et al. is that the reaction energy for the total fragmentation of $A_1-A_2-A_3$ (ΔE_{F1-2-3}) is estimated as the sum of the reaction energies corresponding to the two single-fragmentation processes (i.e., $\Delta E_{F1-2} + \Delta E_{F2-3}$). The straightforward consequence of this approximation is that the energy of the total system can be expressed as a combination of the energies of the three smaller subsystems

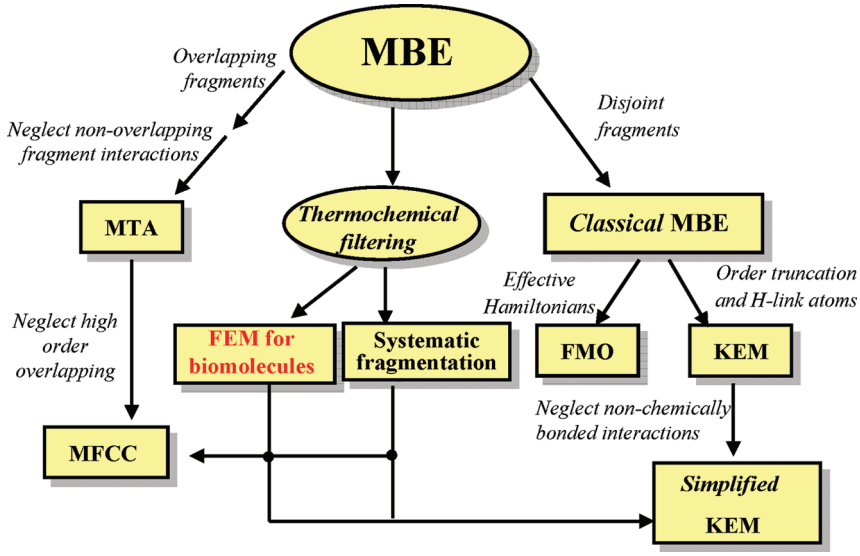
$$E_{123} = E_{12} + E_{23} - E_2 \quad (12)$$

In principle, Collins et al. employ both chemical topology and computer cost considerations in order to choose the best site at which a large molecule is cut so that the resulting A_2 fragment is (a) large enough to reasonably neglect the interaction between A_1 and A_3 and (b) simultaneously small enough to compute the energy of the A_1-A_2H fragment using high-level QM methods. If the accompanying HA_2-A_3 fragment is too large, the fragmentation protocol defined in Scheme 1 is then applied iteratively until all the produced fragments can be described quantum mechanically. Ultimately, this thermochemical approach results in the total energy being approximated by a linear combination of fragment energies, whose precise form depends on the nature of the chemical system and on the chemical topology and computer cost considerations. Like in the MFCC method, the systematic fragmentation technique can be augmented with a nonbonded energy correction by computing the interaction energy between two nonchemically bonded fragments if their separation is below a certain threshold.[25]

Comparison of the Different Methods. Although largely unnoticed in some of the previous works, the MBE formalism provides the general framework for developing computational strategies aimed at the evaluation of the total energy of large systems from subsystem (fragment) energies (see Scheme 2). Thus, the FMO method, the various KEM formulas, and the MFCC expression with pairwise interactions can be classified as MBE techniques that include *N*-body effects through fragment energy calculations. Similarly, the systematic fragmentation method of Collins et al. can be generated directly from the MBE expansion by neglecting all the MBE interaction potentials beyond second order and using an additional chemical topology criterion to neglect a large number of second-order contributions. We can also see in Scheme 2 that inclusion of the H-link atoms to cap the exposed valence sites of the fragments extracted from a covalent system makes the Collins' fragmentation method nearly identical to the simplified version of the KEM method in which only the chemically bonded *double kernels* are considered.[8] Thus, once a fragmentation scheme has been applied, the same energy terms are actually computed in the two methods. Similarly, the systematic fragmentation proposed by Collins et al. encompasses the effective MFCC in which only fragment energies are considered. On the other hand, the MFCC method can be considered as a particular case of the MTA formalism given that the MFCC-capped fragments are equivalent to the MTA overlapping fragments and the MFCC conjugate caps would correspond to fragment intersections in the MTA approach. However, while the MFCC fragments are built to make simple overlaps (i.e., each atom can only be part of one or two fragments), the MTA method admits more complex fragment overlaps among *N* fragments. These and other interrelationships show that in general fragment energy methods assume a similar *ansatz*.

Goals of the Present Work. In principle, the ability to perform on a routine basis fragment energy calculations on large biomolecules could be very useful to predict their energetic properties using high-level QM methodologies. Fortunately, previous test applications have shown that high-order MBE contributions contain many more energetic terms than those that are actually required to derive the total energy from fragment energies within a reasonable accuracy. In this way and taking into account that proteins and nucleic acids are linear polymers that exhibit many repetitive secondary structural motifs, we believe that a thermochemical approach complemented with a distance-based criterion is probably the best option to formulate a linear scaling fragment-based energy method for biological molecules. This approach, which can be considered as a thermochemical truncation of the multibody expansion, is also computationally advantageous because the required energetic terms can be easily computed using standard methodologies. Another advantage of the thermochemical framework is that the successive fragmentation energies involved in the formal degradation of the biomolecule can be computed taking into account the effect of a solvent continuum in the QM Hamiltonian. Thus, in this work, assuming a simple fragmentation process, we will derive a fragment energy formula for estimating the total energy of a biomolecule as function of a cutoff criterion. On one hand, we will show that our fragment energy method (FEM in Scheme 2) can have a broader applicability

Scheme 2



than the MFCC equation, which in turn can be derived from our approach as a particular case. On the other hand, with respect to the more general thermochemical scheme of Collins et al., our expression will be more readily applicable to (and limited to) large biomolecular systems in which a natural choice for the formal fragmentation processes can be easily made. In addition, more emphasis will be placed upon the consistent use of a cutoff criterion in the fragment energy calculations, the inclusion of solvent effects, the mixing of QM and molecular mechanical calculations, and the potential implementation of the fragment-based energy methods within the context of QM/MM methodologies.

Theory

For the sake of simplicity, we will consider a macromolecule ***P*** that is a linear chain of M fragments A_i interconnected through covalent bonds ($A_1-A_2-\ldots-A_M$). For example, if ***P*** is a protein, A_i could be a single amino acid or a secondary structure element. We do note, however, that the same equations based on fragment energies would result for more complex topological patterns connecting the A_i fragments like in cyclic or branched macromolecules.

The total fragmentation of ***P*** can be achieved through the following formal reaction

$$A_1 - A_2 - \ldots A_M + \sum_{i=1}^{M-1}(X_i - Y_i) \rightarrow A_1 - X_1 + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - X_i) + Y_{M-1} - A_M \quad \text{(R1)}$$

Note that every fragment linkage in the ***P*** molecule is broken through insertion of a specific X_i-Y_i molecule(s) into the A_i-A_{i+1} bond. If ***P*** is not a linear chain, then X_i and Y_i would stand for all the molecular caps that are required to saturate the exposed bonds after having removed the A_i fragment from the rest of the ***P*** molecule. In any case, the total energy change corresponding to the above formal reaction is

$$\Delta E = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + E(Y_{M-1} - A_M) - \sum_{i=1}^{M-1} E(X_i - Y_i) - E(\mathbf{P}) \quad (13)$$

The thermochemical approximation to compute ΔE can be introduced as follows: we compute first the reaction energy for the fragmentation step in which the A_1 fragment is removed. However, we assume that the reactants involved in the first fragmentation process are subsystems of ***P*** that are defined on the basis of some geometric and/or chemical-structure criterion. The same criterion, denoted onward as the R criterion, should be applied consistently along the ***P*** backbone structure. Perhaps the simplest criterion for defining the reactants could be to impose a layer cutoff around the leaving A_1 fragment, but other choices like sequence proximity could be used. Thus, assuming that a well-defined R criterion is used, the first fragmentation reaction can be written as

$$A_1 - B_1^R + X_1 - Y_1 \rightarrow A_1 - X_1 + Y_1 - B_1^R \quad \text{(R2)}$$

where B_1^R represents a buffer region, which includes all the neighboring atoms (or fragments A_i) that are around A_1 in the ***P*** structure depending on the R criterion being used. Similarly, the fragmentation process for the A_i-A_{i+1} bond can be represented by the following chemical equation

$$Y_{i-1} - A_i - B_i^R + X_i - Y_i \rightarrow Y_{i-1} - A_i - X_i + Y_i - B_i^R \quad \text{(R3)}$$

where the closer atoms or fragments around A_i excepting those in A_{i-1}, A_{i-2}, ..., A_1 are included in the buffer B_i^R. The sum of the $M-1$ fragmentation processes defined in this manner leads to the following chemical equation

$$A_1 - B_1^R + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1}(X_i - Y_i) \rightarrow A_1 - X_1 + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1}(Y_i - B_i^R) \quad \text{(R4)}$$

In this way, the energy change for the total fragmentation of $\boldsymbol{P}$ through the R-dependent fragmentation processes (ΔE^R) is given by

$$\Delta E^R = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1} E(Y_i - B_i^R) - [E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1} E(X_i - Y_i)] \quad (14)$$

Extracting the exact fragmentation energy ΔE from eq 13 and defining $\delta E = \Delta E^R - \Delta E$, we can combine eqs 13 and 14 in order to exactly express the total energy of the system $E(\boldsymbol{P})$ in terms of the fragment energies and the δE difference

$$E(P) = \left[E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + E(Y_{M-1} - A_M)\right] - \left[\sum_{i=1}^{M-1} E(Y_i - B_i^R)\right] + \delta E(\mathbf{B}^R, \mathbf{Y}) \quad (15)$$

where the δE difference is expressed as a function of $\mathbf{B}^R = \{B_i^R\}$ and $\mathbf{Y} = \{Y_i\}$. This is a consequence of the fact that $E(\mathbf{P})$ is rigorously independent of $\mathbf{B}^R$, $\mathbf{X} = \{X_i\}$, and $\mathbf{Y}$ and that the terms in the square brackets are independent of $\mathbf{X}$ (i.e., the identity of the X_i moieties is irrelevant).

For practical applications of the thermochemical fragment energy eq 15, the δE term must be neglected. To increase the accuracy of the fragment-based energy calculations, one straightforward solution would be to systematically increase the R criterion in order to include larger portions of the remaining $\boldsymbol{P}$ molecule in the B_i^R buffer regions until reaching a reasonable compromise between accuracy and computational cost. The best systems for which we can efficiently apply this simple strategy would be *linear* structures like carbon nanotubes, DNA segments, collagen molecules, etc. Of course, in the case of more compact systems like globular proteins, a larger computational cost and a lower accuracy can be expected for the same R criterion because the buffer regions would contain many more atoms and truncation effects would be more important. However, we could also use the well-known QM/MM methodologies in order to calculate the reaction energies of the fragmentation steps using the same settings as those that are typically employed in routine QM/MM calculations. In this case, the R criterion would be applied to select the size of the QM region while the rest of the system would be treated classically. Thus, like in the electrostatically embedded variants of the MBE methodologies, we expect that QM/MM calculations of fragmentation energies could account for high-order effects within the thermochemical approach.

As above mentioned, we can particularize the general eq 15 to obtain the MFCC equation for a protein system. This can be done by matching Y_i by $-NH_2$ and B_i^R by $-R_{i+1}C_\alpha H_2$, which are the "conjugate caps" adopted in the MFCC scheme. In our thermochemical terminology, these choices are equivalent to consider X_i-NH_2 as the capping dimers as well as to adopt a minimum sequence proximity R criterion for defining the B_i^R groups. Then eq 15 becomes

$$E(\mathbf{P}) = [E(A_1 - C_\alpha H_2 R_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - C_\alpha H_2 R_{i+1}) + E(NH_2 - A_M)] - [\sum_{i=1}^{M-1} E(NH_2 - C_\alpha H_2 R_{i+1})] + \delta E \quad (16)$$

If we compare this equation with eq 11, we see that the "non-neighboring interactions" ($\delta E^{(2)}$) in the MFCC approach[23] constitutes an approximation to the actual error (δE) committed in the calculation of the global fragmentation energy. We note in passing that the same energy contributions collected in eq 16 can be associated to other formal fragmentation processes by changing accordingly the definition of the A_i fragments and the corresponding conjugated caps. For example, expression 16 also results if the A_i fragment corresponds to the i residue and $Y_i = H$.

Finally, it may be interesting to note that our approach, like with all the MBE-like methods, computes the total energy as a linear combination of fragment energies. As gradient is a linear operator, its application over the fragment energy expression would be straightforward as previously noticed in other works.[20,25] In this way, both energy and gradient values for the total system could be obtained from fragment calculations using similar approximations and techniques as those typically used by the QM/MM methodologies.[26,27]

Results and Discussion

In many of the previous works, the viability of fragment-based energy methods has been assessed by means of proof of principle applications, that is, by carrying out single-point calculations and using relatively low QM levels of theory. However, most of the biomolecules are flexible molecular systems in aqueous solution, and therefore, in actual applications, structures for performing fragment-based QM calculations should be provided by Monte Carlo or molecular dynamics (MD) simulations using either explicit or implicit solvent models. In this respect, we think that classical MD simulations still constitute the most reasonable alternative to generate the biomolecular structures for the subsequent fragment QM calculations. This approach would be similar to the molecular mechanics Poisson−Boltzmann method,[28] which predicts mean values of free energies of biomolecules in solution as estimated over a series of representative snapshots extracted from classical MD simulations. Moreover, we also note that various levels of approximation could be required in the fragment energy calculations. For example, a standard density functional level of theory combined with an implicit solvent model can take into account both the intramolecular electronic effects and the solute−solvent electrostatic interactions. Other free-energy terms such as attractive dispersion interactions or thermal contributions could be calculated using molecular mechanics (MM). We believe that this and other technical issues like the counterpoise correction of the basis set superposition error (BSSE) in the QM calculations should be explicitly considered in the test calculations in order to assess the actual performance of the fragment QM energy calculations in the context of multimethod approaches to simulating biomolecules. There-

Scheme 3

$$\begin{array}{lcccc} \alpha\text{–chain} & \text{Pro}-\text{Hyp}-\text{Gly}-\text{Pro}- & (\text{Hyp}-\text{Gly}-\text{Pro} & -\text{Hyp}-\text{Gly} \\ \beta\text{–chain} & \text{Pro}-\text{Hyp}-\text{Gly}- & \text{Pro}-\text{Hyp}-\text{Gly} & -\text{Pro}-\text{Hyp}-\text{Gly} \\ \gamma\text{–chain} & \text{Pro}-\text{Hyp}- & \text{Gly}-\text{Pro}-\text{Hyp})_{i=2,9} & -\text{Gly-Pro}-\text{Hyp}-\text{Gly} \\ & \underbrace{\qquad\qquad}_{A_1} & \underbrace{\qquad\qquad\qquad}_{A_i} & \underbrace{\qquad\qquad}_{A_{10}} \end{array}$$

fore, we decided to reexamine in this work the problem of the stability of triple-helical collagen model peptides by combining our fragment energy expression with previous MD and MM data that have been reported by us recently.[29]

Many collagen model peptides with 30−45 amino acids have been synthesized to investigate the thermal stability and folding of the triple-helix domain of natural collagen. These peptides, which are also known as triple-helical peptides (THPs), assemble spontaneously to form a triple-helix complex that can be characterized using a wide array of experimental techniques.[30] The THP molecules present a characteristic triple-helix structure composed of three peptide chains, each in an extended, left-handed polyproline II-like helix, which are staggered by one residue and then supercoiled about a common axis in a right-handed manner. The close packing of the three chains requires the presence of a sterically small glycine residue at every third position. The test calculations reported in this work were performed on the prototypical $[(\text{Pro-Hyp-Gly})_{10}]_3$ system (labeled as **POG10**), which contains many proline and 4(*R*)-hydroxiproline (Hyp) residues that largely stabilize the triple-helix conformation.[31,32]

Selection of a Fragmentation Process. The collagen model for our test calculations, **POG10**, contains three peptide chains (labeled α, β, and γ) with 30 amino acids per chain. As mentioned above, the fragment energy expression, eq 15, that has been derived by assuming that the ***P*** macromolecule is a linear chain, is also applicable for more complex macromolecules like **POG10**. To this end, we describe the triple helix as a *linear* arrangement of 10 fragments comprising each of three *triplets* of residues from the α, β, and γ chains (see Scheme 3). The resulting building blocks or fragments A_i will be termed as *triplets*. A pair of consecutive *triplets*, A_i-A_j, is interconnected through three peptide linkages corresponding to the α, β, and γ chains. We chose this mode of partitioning because it minimizes the interactions between nonconsecutive *triplets* and maximizes the number of interactions among the three peptide chains within each *triplet*.

After having chosen a structurally and computationally convenient partitioning of **POG10**, we can define more precisely the formal fragmentation processes required for the fragment-energy calculations based on eq 15. More specifically, we see in Figure 1 how the terminating Y_i group attached to the *N*-terminal end of the A_i triplet comprises three acetyl groups for the α, β, and γ peptide chains, whose coordinates are extracted from the *C* end of the previous A_{i-1} triplet and augmented with the required H-link atoms. Similarly, the buffer group B_i^R attached to the *C* end of the A_i triplet includes the adjacent A_{i+1} triplet plus three *N*-methyl moieties extracted from the A_{i+2} fragment (this choice of B_i^R is equivalent to a ∼9 Å cutoff around the leaving A_i fragment). This formal fragmentation process can also be applied straightforwardly to obtain the energy of the individual peptide chains α, β, and γ. In this case, the corresponding A_i, B_i^R, and Y_i groups include residues located in the same chain.

Comparison between Conventional and Fragment-Based QM Energies. Before computing the energy of the full **POG10** system, we assessed the combined quality of the fragment energy calculations and the collagen partitioning in order to reproduce the energetic properties of a relatively large collagen subsystem. The size of the selected subsystem, $[\text{Ace-(Pro-Hyp-Gly)}_4\text{-Nme}]_3$ (456 atoms), still allowed us to carry out full QM calculations. Following similar prescriptions to those represented in Figure 1, four different fragments (A_i) can be distinguished in this model. We computed both the interaction energy among the three peptide chains and the absolute energy of the THP model. The calculations were performed on 25 structures that were built using the coordinates of the central region of POG10 extracted from MD snapshots (see Table S1 in the Supporting Information).[29] As described in the Computational Section, the energy calculations were carried out using a density functional level of theory (PBE/SVP) combined with the COSMO solvent model. The intramolecular dispersion energy is included via an empirical method. The BSSE arising from the interchain interactions is corrected using the standard counterpoise (CP) method. In the case of the fragment energy calculations, the CP correction was applied to the fragment electronic energies, that is, the electronic energies of the $A_1-B_1^R$, $Y_1-A_2-B_2^R$, ..., fragments extracted from one peptide chain (e.g., α) were computed in the presence of the ghost basis functions located in the equivalent fragments from the other two chains (e.g., β and γ). For the full QM calculations, the CP recipe was used to correct the BSSE of the electronic energies of the full peptide chains.

The total interaction energy of $[\text{Ace-(Pro-Hyp-Gly)}_4\text{-Nme}]_3$ can be estimated from the combination of five energy terms using eq 15 (see Table 1). Similarly, the energy of each Ace-(Pro-Hyp-Gly)$_4$-Nme peptide chain can be computed from the corresponding fragment energies. In this way, we derived an average interaction energy (ΔE_{int}) of -29.4 ± 0.2 kcal/mol that matches perfectly the *exact* value (-29.5 ± 0.2 kcal/mol) according to conventional QM calculations.

Since ΔE_{int} is a relative quantity, it can be expected that the fragment energy calculations would benefit from partial cancelation of errors. However, we see in Table 1 that the total energy *E* of the whole system in aqueous solution can be computed accurately using the fragment energies given that the error in the mean value of the fragment-based energies with respect to the exact full QM value is rather small, 0.0001 au (∼0.1 kcal/mol). Table S1 (Supporting Information) shows that small errors arise also in each of the individual structures considered in the calculations. We also see in Table 1 that the observed accuracy in the total energy benefits from a partial cancelation of errors in the computation of the individual energetic components, which result in energy differences of +1.4 (gas-phase energy) and −1.5 kcal/mol (solvation energy) between the fragment-based and the exact values. Although the accuracy in the gas-phase

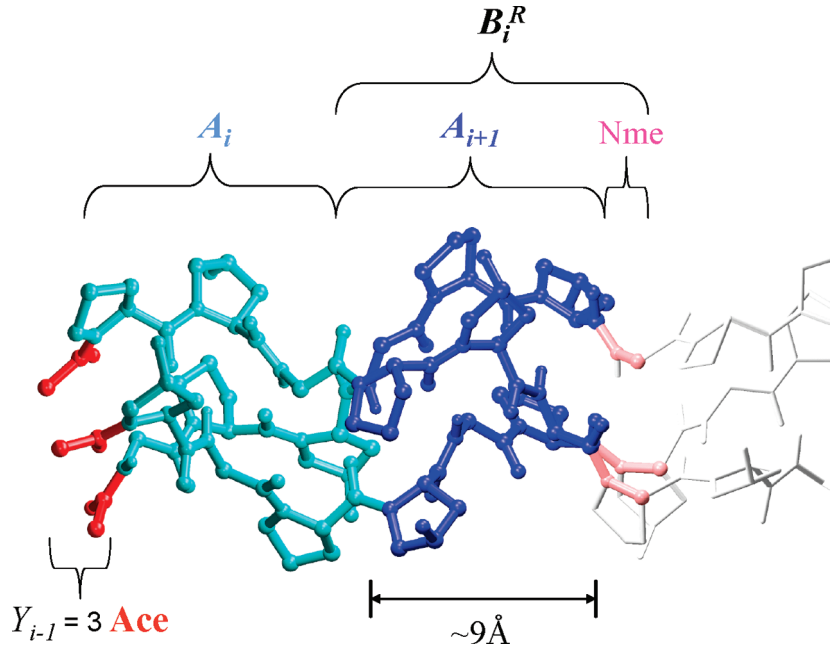


Figure 1. Ball and stick model of the **POG10** triple helix. The various moieties of **POG10** involved in the formal *i*-fragmentation step ($i \geq 2$) are shown in different colors. See text for details.

Table 1. Average Values and Standard Deviations of the Interchain Interaction Energies (ΔE_{int}, in kcal/mol of peptide) for the [Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ System[a]

	$A_1-B_1^R$	$Y_1-A_2-B_2^R$	$Y_2-A_3-B_3^R$	$Y_1-B_1^R$	$Y_2-B_2^R$	[Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ FRAG	[Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ CONV	$\Delta_{FRAG-CONV}$
$\Delta\bar{E}_{int}$	−14.9 ± 0.1	−15.1 ± 0.1	−15.2 ± 0.1	−7.8 ± 0.1	−7.9 ± 0.1	−29.4 ± 0.2	−29.5 ± 0.2	0.1
$\bar{E}$	−6329.7247 (0.0025)	−6329.7250 (0.0021)	−6329.7254 (0.0021)	−3536.8993 (0.0016)	−3536.8983 (0.0014)	−11915.3775 (0.0032)	−11915.3776 (0.0032)	0.1
$\bar{E}_{gas}$	−6329.5131 (0.0023)	−6329.5137 (0.0023)	−6329.5131 (0.0022)	−3536.7781 (0.0015)	−3536.7769 (0.0015)	−11914.9849 (0.0032)	−11914.9872 (0.0033)	1.4
$\Delta\bar{G}^{elec}_{COSMO}$	−87.1 (0.2)	−87.2 (0.3)	−87.7 (0.3)	−54.8 (0.2)	−55.4 (0.3)	−151.9 (0.4)	−150.4 (0.4)	−1.5
$\bar{E}_{disp}$	−105.3 (0.3)	−105.4 (0.4)	−105.3 (0.3)	−53.7 (0.2)	−53.8 (0.2)	−208.4 (0.5)	−208.5 (0.5)	0.1

[a] Average values and standard errors (in parentheses) of the various energy components for the THP fragments: total energy in solution, E, in au; gas-phase energy, E_{gas}, in au; electrostatic solvation energy, ΔG^{elec}_{COSMO}, in kcal/mol; and empirical dispersion energy, E_{dis}, in kcal/mol. Mean values of the total energies as obtained with the fragment-based (FRAG) and conventional (CONV) calculations and their differences ($\Delta_{FRAG-CONV}$, in kcal/mol) are also indicated.

Table 2. Average Values (kcal/mol of peptide) for the Different Energy Components of the Interaction Energy among the **POG10** Peptide Chains[a]

$\Delta\bar{E}^{CP\text{-}uncorrected}_{PBE/SVP}$	BSSE	$\Delta\bar{G}^{elec}_{COSMO}$	$\Delta\bar{E}_{disp}$	$\Delta\bar{E}_{int}$[b]
−105.6 (1.1)	85.7 (0.2)	37.1 (1.0)	−82.5 (0.1)	−65.4 (0.2)

[a] Standard errors are given in parentheses. [b] $\Delta\bar{E}_{int} = \Delta\bar{E}^{CP\text{-}uncorrected}_{PBE/SVP} + BSSE + \Delta\bar{G}^{elec}_{COSMO} + \Delta\bar{E}_{disp}$.

energy (~0.002 au) is comparable to that reported in previous fragment energy calculations,[10,14,20] these results suggest that inclusion of solvent effects in the fragment QM calculations should improve the accuracy of the fragment-based approaches given that the intramolecular long-range interactions could be dampened out by the electrostatic screening exerted by the surrounding solvent continuum.

Due to the linear structure of collagen, we expect that the performance of the fragment-energy calculations for larger collagen models would be equally satisfactory and that other molecular properties of collagen molecules (e.g., gradients) could be also computed within a reasonable accuracy. Finally, we note that, in terms of CPU time, a single-point energy calculation on the [Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ system using the fragment approach took about 9 h on one x86−64 processor. The same energy value obtained with conventional QM calculations required about 80 h of CPU time.

Fragment Calculations on the POG10 Triple Helix. The results of our fragment energy calculations on the full **POG10** system (1089 atoms) are summarized in Table 2, which contains the average values of the various energetic components contributing to the interchain interaction energy. The calculations were done on 100 snapshots extracted from our previous MD simulation.[29] The total interaction energy amounts to −65.4 kcal/mol of peptide, which gives an average value of −6.5 kcal/mol for every −(Pro-Hyp-Gly)− triplet of residues. As expected, all the energy components considered in the calculations (gas-phase electronic energy, empirical dispersion energy, and electrostatic solvation

energy) contribute significantly to the interaction energy. Of particular interest can be the large weight of the BSSE as estimated by the CP calculations, 85.7 kcal/mol. Clearly, the omission of the BSSE corrections would have resulted in an unphysical overestimation of the interaction energy. On the other hand, the inability of the PBE DFT functional to recover most of the intermolecular dispersion energy justifies the addition of the empirical dispersion energy. In fact, the combination of DFT QM methods and empirical dispersion energy has been used in previous computational studies that apply DFT to study weak nonpolar interactions.[33−35] Although the three peptide chains intertwined into the triple helix establish many hydrogen-bond interactions that can be described reasonably by the PBE calculations, we see in Table 2 that the dispersion energy is the largest stabilizing contribution to the interchain interaction energy of the **POG10** triple helix. Hence, it turns out that the close packing of the peptide chains plays a crucial role in the overall stabilization of the triple helix.

Perhaps the bottom line from the calculations summarized in Table 2 is that the QM fragment energy approach may constitute a promising alternative for studying the intermolecular interactions in large biomolecules. For the collagen model peptide studied in this work, the error introduced by the fragmentation technique can be rather small (<1 kcal/mol) as suggested by the preliminary test calculations. However, we do note again that when using a DFT level of theory in the fragment calculations for large biomolecules, correction of the BSSE and inclusion of dispersion energy are a must in order to obtain meaningful results for interaction energies.

Intramolecular BSSE. As shown in Table 2, the CP correction to the interchain interaction energy is quite large, +85.7 kcal/mol at the PVE/SVP level, due to the large size of the **POG10** system and the relatively small size of the double-ζ SVP basis set. In principle, the use of larger basis sets should reduce significantly the magnitude of the BSSE but at the cost of increasing the CPU time. Nevertheless, it is most likely that assessing and correcting the BSSE will also be required when carrying out fragment energy calculations on biomolecules using medium-sized basis sets (cc-pVDZ, TZVP, ...). Moreover, it is becoming increasingly clear that the relative energies of different conformations of large and flexible biomolecules are quite sensitive to the size of the basis set and that part of this dependence arises from the *intramolecular* BSSE.[36] Although this (presumably small) effect has been commonly ignored so far, there is now some solid computational evidence in the recent literature indicating that the intramolecular BSSE can severely impair the accuracy of the energetic QM predictions for polypeptide systems.[36−38]

Given that we are interested in computing the relative stability of the triple-helix conformation with respect to the compact form of the isolated chains (see below), we decided to estimate the magnitude of the intramolecular BSSE in our QM calculations. For this purpose, the CP method of Boys and Bernardi could be applied by taking atomic fragments, but this alternative would result in a large number of extra QM calculations as well as in problems in the assignation of charge, multiplicity, and electronic state of the atomic fragments.[39] Hence, we followed a more pragmatic approach that consists of the definition of proper molecular fragments within the large system and adding H-link atoms to saturate the exposed chemical bonds. Subsequently, the BSSE in the interaction among the resulting fragments is computed using the standard CP procedure. A similar approach has been employed previously by other authors.[36] For example, Valdés et al. estimated the intramolecular BSSE in [*n*]-helicene molecules consisting of all-ortho-annulated benzene rings by computing the CP-corrected interaction energies of benzene pairs, in which the Cartesian coordinates of the C atoms are identical to those in the helicene.[36]

After some computational experimentation, we decided to employ the following fragmentation protocol for estimating the intramolecular BSSE of the **POG10** peptide chains. (1) For each **POG10** structure, a pair list of nonbonded (beyond 1−4) interactions involving heavy atoms is built using a distance criteria (X···Y < 4.0 Å). (2) Each peptide chain is broken into four smaller fragments by removing three glycine residues. These glycine residues are automatically selected in order to *maximize* the number of nonbonded interactions among the resulting fragments (see Figure 2a and 2b). H-Link atoms are added to saturate the exposed bonds. (3) The standard CP method is used to compute the value of the BSSE corresponding to the interactions among the four fragments (*intra*-$BSSE_1$; see Figure 2b). (4) The BSSE due to the interactions between the formerly removed glycine residues and the nearby groups is estimated by building a molecular cluster in which the three glycine residues are surrounded by the closer residues. Then the CP procedure is applied again to estimate the BSSE arising from the simultaneous interactions between the three glycines and the rest of the groups (*intra*-$BSSE_2$; see Figure 2c). (5) The total intramolecular BSSE of the peptide chain is approximated by adding together the two BSSE values computed in 3 and 4.

The QM calculations for estimating the intramolecular BSSE were done on 100 MD snapshots of the free **POG10** chain.[29] Thus, we found that, at the PBE/SVP level, the average value of the intramolecular BSSE for the isolated **POG10** chain in its folded state amounts to 92.7 kcal/mol of peptide, which is even greater than the BSSE related to the interchain interactions in the triple-helix state (85.7 kcal/mol). For the sake of consistency, the same protocol was applied on each of the three chains in the triple-helix conformation. In this case, the peptide chains are quite extended and their intramolecular BSSE is predicted to be only 3.1 kcal/mol on average. All these CP-corrected QM calculations can be combined to estimate the energetic penalty for the folded **POG10** chain to adopt its extended conformation in the triple helix, the average value being +30.8 kcal/mol (in terms of $E_{PBE/SVP} + BSSE_{intra} + E_{disp} + \Delta G^{elec}_{COSMO}$). Neglecting the intramolecular BSSE in the folded state of **POG10** would lead to a very large unrealistic value (∼120 kcal/mol) for the relative energy between the folded and the extended forms of the peptide chain.

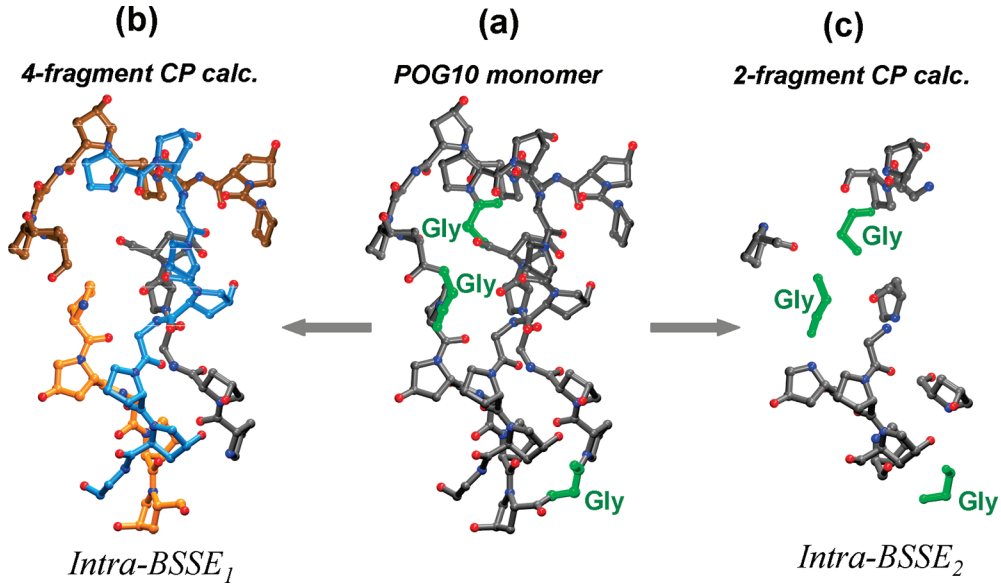


Figure 2. Ball-and-stick models of a **POG10** chain in its monomer state showing the fragmentation procedure followed to correct the *intramolecular* BSSE through CP calculations. (a) On the basis of a nonbonded interaction pair list, three glycine residues (in green) are selected in order to maximize the number of nonbonded interactions among the peptide fragments that result upon removal of the glycine residues. (b) BSSE arising from the interactions among the four peptide chains (C atoms are shown in different colors) is estimated using the CP procedure. (c) A molecular cluster is constructed from the coordinates of the glycine residues selected in a and those of the nearby peptide residues that interact directly with the marked glycines. The BSSE associated to the interaction between the glycines and the nearby groups is again estimated by means of CP calculations.

Table 3. Average Values and Standard Errors (in kcal/mol of peptide) of the Free-Energy Components for the Transition from the Monomeric to the Triple-Helix State at 300 K

	mean value	standard error		mean value	standard error
$\Delta\bar{E}_{PBE/SVP}$	53.7	*9.7*	$\Delta\bar{H}^{norm}_{MM\text{-}GBSA}$	0.7	*0.1*
$\Delta\bar{E}^{CP\text{-}corrected}_{PBE/SVP}$	49.8	*9.6*	$-T\Delta\bar{S}^{norm}_{MM\text{-}GBSA}$	8.8	*0.9*
$\Delta\Delta\bar{G}^{elec}_{COSMO}$	−73.2	*9.3*	$-T\Delta S_{conf}$	0.4	-
$\Delta\bar{E}^{solute}_{disp}$	−10.7	*0.4*	$\Delta\bar{G}^{CP\text{-}corrected\ a}$	−11.7	*1.8*
$\Delta\bar{E}^{solute-solvent}_{disp}$	11.0	*0.8*	$\Delta\bar{G}^{a}$	−7.8	*2.1*
$\Delta\bar{G}_{cav}$	−1.2	*0.1*			

[a] Assuming a standard state of 0.001 M.

Free Energy for the Transition from Monomer to Triple Helix. As shown above, the fragment QM calculations complemented with the empirical dispersion formula can give insight into the nature of the interactions holding the peptide chains in the triple-helix conformation. However, the actual stability of the triple helix is determined by the free-energy change for dissociation to give the free peptide monomers. In our previous work,[29] we found that the isolated **POG10** peptide in aqueous solution adopts a stable folded conformation, and therefore, by combining the fragment QM data on the triple helix with the results of QM calculations on a representative set of **POG10** monomers, one could estimate the corresponding free-energy change for the peptide aggregation process leading to the **POG10** triple helix, provided that the selected QM method gives a compensated description of the conformational and intermolecular interaction energies. By taking advantage of our previous computational experience, we combined the QM energies with further molecular-mechanical data in order to ensure a balanced description of other free-energy components (solute−solvent vdW interactions, thermal contributions to free energy, etc.). More specifically, we used the following expression in order to compute the average free energy of the **POG10** system both in its triple-helix and monomer states

$$\bar{G} = \bar{E}^{CP-corrected}_{PBE/SVP} + \bar{E}^{solute}_{disp} + \bar{E}^{solute-solvent}_{disp} + H^{norm}_{MM-GBSA} - T\bar{S}^{norm}_{MM-GBSA} + \Delta\bar{G}^{elec}_{COSMO} \quad (17)$$

where the gas-phase $\bar{E}^{CP\text{-}corrected}_{PBE/SVP}$ energy, which includes the intermolecular and intramolecular BSSE corrections, and the electrostatic solvation energy ($\Delta\bar{G}^{elec}_{COSMO}$) are computed by means of fragment-based (triple helix) and standard (monomer) QM calculations; the $\bar{E}^{solute}_{disp} + \bar{E}^{solute-solvent}_{disp}$ dispersion energy terms are computed with the same empirical formula, and normal mode molecular mechanical calculations are used to estimate the thermal contributions to free energy. The change in the average values of these energetic components for the monomer → triple-helix transition are collected in Table 3, which also includes the corresponding small differences in the cavitation free energy and the conformational entropy that were computed following the procedures described in our previous work.[29]

We see in Table 3 that the QM energy terms (gas-phase and solvation energy) as well as the empirical dispersion

energies change significantly on going from the monomer to the triple helix. In agreement with our previous molecular mechanical and Poisson−Boltzmann (MM-PB) calculations, the QM-based approach predicts also that the driving force for the formation of the triple helix is mainly provided by the electrostatic solvation energy. The total ΔG value obtained with the CP-corrected QM energies amounts to −11.7 kcal/mol, with a statistical uncertainty of 1.8 kcal/mol (standard error). This value is in moderate agreement with the most accurate experimental estimate at 300 K, −6.4 kcal/mol, which has been derived from differential scanning calorimetry.[29,40] The purely MM-PB calculations together with a broader sampling give a ΔG value of −6.2 (1.2) kcal/mol.[29] The larger difference between the QM-based calculations and experiment is most likely due to several factors like the small error in the fragment-based QM calculations, the remaining inaccuracy in the correction of the intramolecular BSSE, slight unbalances in the combination of QM and MM data in eq 17, as well as by some limitations of the PBE DFT functional to reproduce the electrostatic and H-bond interactions. All these potential sources of error, which are not present in the MM-PB calculations, could be mitigated by gaining more computational experience and improving the details of the mixed QM−MM computational protocol. On the other hand, it turns out that the ΔG value obtained with the CP-uncorrected QM energies (−7.8 kcal/mol) is closer to the experimental estimate. Nevertheless, this result is somewhat fortuitous given that, in the particular case of the **POG10** system, the sum of intra- and intermolecular interactions remains approximately constant upon the monomer → triple-helix transition.

Summary and Conclusions

In this work we reviewed several computational methods developed during the last years for computing the energy of large molecules using only fragment energies. Although some of the previous methods have been introduced independently to each other, a comparative analysis reveals their common roots, which, in our opinion, can be traced back to the general formalism of the MBE method. For biomolecules constructed with repetitive building blocks (residues, secondary structural elements,...), it is proposed that a simple thermochemical approach is probably the best option for formulating a *standard* fragment energy method. The validity of the fragment QM energy strategy has been tested intensively considering a challenging problem for simulation methodologies, that is, the prediction of the interchain interaction energy and the free energy for dissociation of a prototypical collagen model. The comparison of our fragment-based energies with experimental data and former theoretical results shows that the actual applicability of the fragment QM methods in biomolecular simulations will rely heavily on the proper combination of QM and MM calculations as well as in the conformational sampling performed by MM methods. Moreover, the correction of the inter- and intramolecular BSSE will be critically important for obtaining realistic energies of either interaction or conformational changes.

Since the MM-PB method predicts a more accurate value than the fragment-based QM calculations for the ΔG change in the monomer → triple-helix transition of the **POG10** system, one may raise the question of whether the fragment QM approaches are really needed. Clearly, the fragment QM calculations would have a broader applicability since they can be used to investigate all kinds of interactions and chemical transformations involving biomolecules. For example, most of the current force fields have been developed without specifically considering the interactions of biomolecules with metal ions, clusters, or surfaces, and therefore, the application of fragment-QM methodologies to study *biomaterials* could provide reliable energetic data, which in turn could be useful for the development and validation of new MM parameters. In addition, we point out that the QM charge densities obtained in the fragment calculations contain much valuable information that can be used for estimating other QM properties (e.g., electrostatic potential) and deriving QM descriptors (e.g., for determining ligand affinity). Similarly, the fragment QM calculations could also be used to outline electron pathways connecting the electron donor and acceptor sites in redox metalloproteins[41] and the energy gaps between electronic states. Therefore, with the continuous improving in the efficiency of QM methodologies, the decreasing cost of computer hardware, as well as a necessary standardization of the fragment energy approach by means of intensive computational experimentation, the full QM description of large biomolecules could be done regularly in the near future.

Computational Methods

DFT Calculations. Density functional theory methods have become the most popular QM methodology for the study of biomolecules because they include electron correlation effects at a relatively cheap computational cost. In principle, the Perdew−Burke−Ernzerhoff (PBE)[42] and Tao−Perdew−Staroverov−Scuseria (TPSS)[43] functionals are particularly attractive for performing fragment energy calculations, since they are nonempirical GGA functionals that give results with an acceptable quality in any type of chemical systems including macromolecules and condensed phases. In this work, we used the PBE functional combined with a double-ζ plus polarization basis set (SVP). [44] The reliability of the PBE/SVP level of theory was assessed by carrying out some validation calculations on a small triple-helix system (see below).

All DFT calculations were performed using the TURBOMOLE suite of programs,[45] in the framework of the multipole accelerated resolution-of-the-identity approximation (MARI-J) using the appropriate auxiliary basis set.[46,47] To estimate the effect of the solvent environment on the DFT energies, we used the conductor-like screening model (COSMO) included in TURBOMOLE in which the solvent dielectric continuum is approximated by a scaled conductor.[48] The optimized atomic COSMO radii (r_H = 1.3 Å, r_C = 2.0 Å, r_N = 1.83 Å, and r_O = 1.72) were used to generate the solvent-accessible molecular cavity.[49] Note that in the thermochemical fragment energy calculations reported in this work long-range electrostatic effects are truncated in the different fragment calculations and that, therefore, a molecular cavity is constructed around each fragment system

Table 4. Average Values and Standard Deviations for the Interaction Energy (kcal/mol of THP) among the Three Peptide Chains for 25 Snapshots of the [Ace(Pro-Hyp-Gly)-Nme]$_3$ Trimer

level of theory	ΔE_{int}	level of theory	ΔE_{int}
PBE/SVP[a]	-10.7 ± 1.4	PBE/SVP[b]	-10.7 ± 6.2
PBE/TZVP[a]	-8.6 ± 1.3	PBE/TZVP[b]	-8.2 ± 5.8
PBE/TZVPP[a]	-9.0 ± 1.3		
TPSS/SVP[a]	-9.0 ± 1.1		

[a] Geometries were extracted from the **POG10** MD simulations and relaxed via MM energy minimization. [b] Geometries were extracted from the **POG10** MD simulations.

($Y_i - B_i^R$, $Y_{i-1} - A_i - B_i^R$, ...). This is fully consistent with the estimation of the full system energy from a combination of reaction energies (eqs R3 and R4).

Since the GGA density functionals are unable to describe dispersive interactions, the DFT energy terms were augmented with an dispersion energy contribution, E_{disp}, which was computed using an empirical formula that has been introduced by Elstner et al.[34] in order to extend their approximate DFT method for the description of dispersive interactions. The E_{disp} expression consists basically of a $-C_6/R^6$ term, which is appropriately damped for short R distances. We used the same parameters for C, N, O, and H and combination rules as those described by Elstner et al.[47]

Molecular Geometries and Molecular Mechanical Calculations. Molecular geometries of the **POG10** system were taken from our previous study on the relative stability of collagen model peptides.[29] The triple-helix and monomer states of **POG10** were subject to 20 and 50 ns molecular dynamics (MD) simulations, respectively, at constant P (1 atm) and T (300 K) in explicit solvent using the AMBER package.[50] From these MD simulations, a set of 100 snapshots was extracted for each state and the internal geometry of the solute molecules was relaxed throughout energy minimization prior to the QM and MM energy calculations. The snapshots were postprocessed through the removal of all solvent molecules.

Thermal contributions to the enthalpy and entropy of solute molecules were estimated by means of MM normal mode calculations using the NAB package[51] and following the prescriptions described elsewhere.[29] The nonpolar solvation energy was computed by combining the explicit solvent representation with an estimation of the relative change in the cavitation free energy of the solute.[52] In our previous work, the conformational entropy of the solute was computed via an expansion of the so-called mutual information functions.[6]

Validation Calculations of the PBE/SVP Level of Theory. Table 4 summarizes the results of some preliminary validation calculations in which we computed the interchain interaction energy in a small THP model ([Ace-(Gly-Pro-Hyp)-Nme]$_3$; 123 atoms). In these calculations, we used the PBE and TPSS functionals combined with different basis sets ranging from the double-ζ SVP to the triple-ζ plus double polarization TZVPP. All DFT energies include the effect of aqueous solvent (COSMO model) and are combined with the empirical estimate of the dispersion energy. We also corrected the BSSE affecting the intermolecular interaction energy by means of the counterpoise method. Coordinates of the small THP models were taken from 25 truncated snapshots of our previous MD simulations of the **POG10** system after having relaxed the internal geometry of the solute molecules via energy minimizations using the AMBER force field.

We see in Table 4 that the average PBE energies obtained with various basis sets are quite similar, the differences being around 1−2 kcal/mol. The TPSS functional gives similar interaction energies to those provided by PBE. By repeating some calculations without relaxing the internal geometry of the small THP models, we found that the average interaction energies are hardly affected, but standard deviations are much higher (~6 kcal/mol). Overall, we conclude that the PBE/SVP energy calculations on the MM-relaxed geometries may constitute a reasonable compromise between quality and computational cost.

Acknowledgment. This research was supported by the following grants: FICyT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. E.S. and N.D. thank MEC for their FPU and Ramon y Cajal contracts, respectively. We are grateful to Dr. H. Valdés for her careful reading of the manuscript and suggestions.

Supporting Information Available: Equivalence between second-order MBE and KEM; Tables S1 and S2 containing the relative and absolute energies of all the MD structures considered in the test calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.

(2) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879.

(3) Connolly, J. W. D.; Williams, A. R. *Phys. Rev. B* **1983**, *27*, 5169.

(4) Carlsson, A. E. Beyond pair potentials in elemental transition metals and semiconductors. In *Solid State Physics*; Ehrenreich, H., Turnbull, D., Eds.; Academic Press: Boston, 1990; Vol. 43, p 1.

(5) Drautz, R.; Fähnle, M.; Sanchez, J. M. *J. Phys.: Condens. Matter* **2004**, *16*, 3843.

(6) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096.

(7) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808.

(8) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 447.

(9) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 12690.

(10) Huang, L.; Massa, L.; Karle, J. *J. Chem. Theory Comput.* **2007**, *3*, 1337.

(11) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2008**, *105*, 1849.

(12) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.

(13) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *20*, 6832.

(14) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(15) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.

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

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

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

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

(20) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.

(21) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484.

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

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

(24) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

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

(26) Vreven, T.; Frisch, M. J.; Kudin, N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701.

(27) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.

(28) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889.

(29) Suarez, E.; Diaz, N.; Suarez, D. *J. Phys. Chem. B* **2008**, *112*, 15248.

(30) Brodsky, B.; Persikov, A. V. *Adv. Protein Chem.* **2005**, *70*, 301.

(31) Bella, J.; Brodsky, B.; Berman, H. M. *Structure* **1995**, *3*, 893.

(32) Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75.

(33) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

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

(35) Jureka, P.; Cerný, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(36) Valdés, H.; Klusák, V.; Pitoák, M.; Exner, O.; Starý, I.; Hobza, P.; L., R. *J. Comput. Chem.* **2008**, *29*, 861.

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

(38) Palermo, N. Y.; Csontos, J.; Owen, M. C.; Murphy, R. F.; Lovas, S. *J. Comput. Chem.* **2007**, *28*, 1208.

(39) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Phys.* **2008**, *128*, 144108.

(40) Nishi, Y.; Uchiyama, S.; Doi, M.; Nishiuchi, Y.; Nakazawa, T.; Ohkubo, T.; Kobayashi, Y. *Biochemistry* **2005**, *44*, 6034.

(41) Guallar, V. *J. Phys. Chem. B* **2008**, *112*, 13460.

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

(43) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

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

(45) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(46) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.

(47) Eichkorn, K.; Treutler, O.; Ohm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(48) Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenzc, J. C. W.; Eckert, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187.

(49) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. *J. Phys. Chem. A* **1998**, *102*, 5074.

(50) Case, D. A.; Darden, T. A.; Cheatham, I., 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, 2006.

(51) Macke, T.; Case, D. A. Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids*; Leontes, N. B., SantaLucia, J. J., Eds.; American Chemical Society: Washington, DC, 1998; pp 379.

(52) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238.

CT8005002

JCTC Journal of Chemical Theory and Computation

Validating CHARMM Parameters and Exploring Charge Distribution Rules in Structure-Based Drug Design

Jennifer L. Knight and Charles L. Brooks, III*

Department of Chemistry and Department of Biophysics, University of Michigan, 930 North University Avenue, Ann Arbor, Michigan 48109

Received February 14, 2009

Abstract: Using an extensive series of TIBO compounds that are non-nucleoside inhibitors of HIV-1 reverse transcriptase, we have systematically evaluated the quality of recently developed ligand parameters that are consistent with the CHARMM22 force field. Thermodynamic integration simulations for 44 pairs of TIBO compounds achieve a high level of success with an overall average unsigned error (AUE) in the relative binding affinities of 1.3 kcal/mol; however, the accuracy is strongly dependent on the size differential between the substituents sampled as well as the class of functional group. Low errors are observed among the alkyl, allyl, aldehyde, nitrile, trifluorinated methyl, and halide TIBO derivatives, and large systematic errors are observed among thioether derivatives. We have also investigated how different charge assignment schemes for small molecules impact the quality of computed binding affinities for a subset of this series. This study demonstrates the advantage of using model compounds to derive physically meaningful charge distributions and bond-charge increments for rapidly expanding fragment libraries for drug development applications. Specifically, in the absence of a bond-charge increment for a given pair of atom types, the strategy of adopting CHELPG charges from localized regions of model compounds provides reliable results when modeling with the CHARMM force field.

Introduction

Computational methods have become important resources in structure-based drug design.[1,2] Three-dimensional structures can be used to model the interactions between protein targets and potential new drugs and to predict their binding free energies.[3−5] Empirical all-atom force fields that are used to represent proteins in these simulations have matured to a significant level.[6,7] However, due to the enormity of chemical space, it is still challenging to develop force field parameters that cover a wide range of compounds that might be encountered in drug design and development efforts.[7]

Ligand parametrization procedures are traditionally computationally intensive and can represent a bottleneck in structure-based drug design strategies. To develop force field parameters that are tailored for a new compound, the specific parameters required for the intra- and intermolecular energy terms must be optimized. This process may require several iterations until the parameters yield appropriate conformational energies, hydration free energy, dipole moment, or other molecular properties of the modeled compound. These ligand parametrization efforts may be accelerated if information about well-parametrized compounds can be leveraged to describe new compounds under investigation. However, individual force fields have been developed with different philosophies, which means that, in general, ligand parameters are not immediately transferable among the biomolecular force fields.[8,9]

Each of the major biomolecular force fields have programs that read in the coordinates of a compound and assign atom types, partial charge distributions, and energy parameters on the basis of information in template libraries.[10−15] For example, the molecular modeling package IMPACT[10] and the utility software script, hetgrp_ffgen (Schrödinger, LCC), as well as the BOSS and MCPRO[16] molecular modeling systems (Cemcomco, LLC) facilitate modeling with the

* Corresponding author phone: (734) 647-6682; fax: (734) 647-1604; e-mail: brookscl@umich.edu.

10.1021/ct900079t CCC: $40.75
Published on Web 05/21/2009

OPLS-AA[17] force field; Antechamber[11] was developed as an auxiliary program in the AMBER[12] molecular modeling packages; PRODRG[13,14] prepares ligands for modeling with the GROMOS force field;[18] and the recently developed MATCH suite of tools (unpublished, D. J. Price and C. L. Brooks, III) constructs ligand files that are compatible with the CHARMM[19] force field. The success of these automated parametrization programs depends on the extent of the classes of compounds that are covered within the template libraries, the quality of the parameters themselves, and the transferability of parameters from the modeled compounds or fragment to a novel context.

Significant progress has been made to develop ligand parameters that are compatible with the CHARMM22 force field and are transferable from smaller model compounds into more complicated chemical structures. Mackerell and co-workers have most recently introduced newly optimized halide and ether parameters to this CHARMM General Force Field (CGenFF) (private communication, K. Vanommeslaeghe and A. D. Mackerell). Often, the quality of force field parameters is assessed by their ability to reproduce the hydration free energies of small molecules or thermodynamic properties of bulk solutions.[20−23] However, the primary end-use of these ligand parameters is to model the interactions between putative drug compounds and larger biomolecules, like proteins and nucleic acids. In this study, we will evaluate a variety of CGenFF parameters for their ability to reproduce relative binding affinities for a series of compounds. To our knowledge, this work represents the first large-scale assessment of the quality of CGenFF parameters in the context of binding free energy calculations.

We have chosen the TIBO class of non-nucleoside inhibitors of HIV-1 reverse transcriptase (RT) because of the availability of extensive experimental data and because it has been used in a variety of contexts[24−26] as a benchmark for evaluating the quality of free energy models using AMBER and OPLS-AA force fields. Specifically, linear interaction energy (LIE) models as well as molecular dynamics simulations coupled with MM-PBSA simulations have achieved high levels of success in computing absolute binding affinities for series of these TIBO-like compounds bound to HIV-1 RT. Smith et al.[25] examined 12 TIBO derivatives using the OPLS-AA force field, and their best linear response approximation models obtained root-mean squared errors of 0.9 kcal/mol, although no test set was included to provide a more unbiased estimate of the uncertainty in the calculations. Wang et al.[24] tested this same set of 12 compounds and with molecular dynamics and MM-PBSA calculations governed by the AMBER force field predicted binding affinities with errors on the order of ∼1 kcal/mol, and the largest error was 1.9 kcal/mol. Su et al.[26] computed binding affinities for 37 TIBO compounds using the OPLS-AA force field and achieved average LIE model errors as low as 1.2 kcal/mol for predicting the binding affinity of one compound given the LIE parameters that were fit to the remaining 36 compounds. The high quality of these results irrespective of method and force field suggests that the HIV-1 RT:TIBO system is relatively well-behaved and thus serves as a good benchmark for evaluating the quality of new ligand parameters. In contrast to these previous studies, we perform a series of thermodynamic integration calculations so that we can readily identify systematic errors that relate to specific classes of compounds and ascertain where improved force field parameters are warranted and so that we do not need to estimate entropy contributions.

Charge Distribution Rules in Structure-Based Drug Design. Arguably, partial charges are the most difficult ligand parameters to transfer among force fields or to adopt from other "known" molecules within a given force field due to their dependence on their local bonded environment. Yet, assigning appropriate charge distributions in novel compounds is of profound importance in effectively representing the nonbonded interactions in binding free energy calculations.[27] Atomic partial charges are the primary components of the electrostatic energy terms and are critical for adequately describing the correct desolvation penalty when a small molecule is transferred from solution into a binding pocket. Certainly polarization effects influence the magnitude of the desolvation penalty as well as the strength of the protein−ligand interaction energy and will play a more significant role in the presence of larger differences in the dielectric properties between the solvent and the binding pocket. While polarizable force fields are being developed,[28,29] most biomolecular force fields rely predominantly on fixed-charge models.

Two main strategies have been suggested for generating partial charge assignments that are compatible with current biomolecular force fields. In one fixed-charge strategy, charges are adopted for an entire molecule, often based on ab initio calculations. For example, a restrained electrostatic potential (RESP) charge fitting procedure or a semiempirical method that mimics these charge distributions is advised for assigning partial charges to novel ligands in a manner that is consistent with the generalized AMBER force field (GAFF).[30,31] Systematic studies using Lennard-Jones parameters from the OPLS or AMBER force fields demonstrated that partial charge distributions that were fit to electrostatic potentials (ESP) or scaled CM1A partial charges yield hydration free energies for small molecules that have average errors on the order of 1 kcal/mol.[21−23] However, several chemical classes, especially the more polar compounds, exhibit larger individual errors. The largest unsigned error observed for solvation free energies modeled by GAFF using semiempirical AM1 charges with Bayly and coworkers' parametrized bond-charge corrections (AM1-BCC) was about 3 kcal/mol,[22] while the results for the OPLS-AA force field with the Cramer/Truhlar CM1A charge model scaled by 1.14 led to maximal errors of about 2.5 kcal/mol.[21]

In the second fixed-charge strategy, generally employed by CHARMM and OPLS-AA force fields, bond-charge increment (BCI) "rules" are employed such that optimal charges are determined for fragments of molecules, and then these fragments are pieced together to construct charge distributions for novel compounds.[32] In addition to variations in the specific force field parameters, these programs differ in how the bonded environment is determined, how the specific BCI rules are defined for matching the fragments in

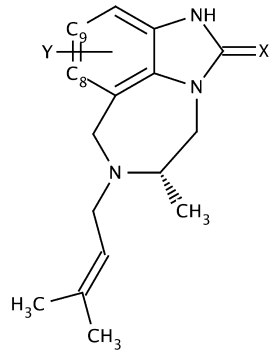


Figure 1. The TIBO core structure.

the new molecule with those "known" fragments, and how excess charges are distributed throughout the molecule.

In developing these libraries of fragments, it is important to ascertain whether more automated processes could effectively determine charge distribution for the fragments themselves and where these optimization strategies could be transferable across a variety of functional groups. In addition, it is imperative that sufficient information about the charge distributions from these well-parametrized fragments be included to adequately describe new compounds under investigation. If too little information about the chemical context of the fragment is included, properties of the subsequent compounds may lack appropriate specificity and binding affinities may be unreliable. On the other hand, if too much specificity about the chemical context is required, then the modeled fragments become less transferable to new compounds. Both of these issues are addressed in this study in attempts to focus subsequent ligand parametrization efforts.

In the present Article, we examine the range of questions discussed above. First, we validate a variety of CGenFF parameters for use in structure-based drug design. Thermodynamic integration (TI) simulations are used to compute the relative binding affinities for select pairs of these 21 TIBO derivatives. These TIBO compounds possess a common chemical core structure and only differ from one another at one of two substituent sites. Therefore, systematically evaluating this series of compounds mimics a chemical optimization strategy in which various substituents or fragments are evaluated at specific sites on a promising new therapeutic lead compound.

Second, we explore the effect of different charge distribution rules in structure-based drug design for constructing new bond-charge increments. Four charge distribution schemes are investigated to determine what features of the charges of the constitutive fragments contribute to the accuracy of the computed binding affinities of the TIBO derivatives. These schemes differ in how the charges are assigned and the extent to which a given fragment influences that charge distribution in the rest of the molecule.

Methods

Ligand Set and Experimental Binding Affinities. Figure 1 and Table 1 show the molecular structures and the experimental binding free energies of the 21 TIBO compounds that were included in these calculations. These ligands and their corresponding IC_{50} values were compiled from Ho et al.[33] and Smith et al.[25] Differences among the compounds are limited to two variations at the X site (C═O and C═S) and 14 variations at the Y site (alkanes, alkynes, halides, trifluorinated methyls, nitriles, aldehydes, ethers, and thioethers) on the TIBO core.

Table 1. Molecular Structures and the Corresponding Experimental IC_{50} and Binding Free Energies of the TIBO Analogues

compound	X	Y[a]	IC_{50}[b] (μM)	$\Delta G_{binding}$[c] (kcal/mol)
1	S	Br	0.0030	−12.09
2	S	Cl	0.0043	−11.87
3	S	SCH_3	0.0050	−11.78
4	S	F	0.0058	−11.69
5	S	CH_3	0.0136	−11.16
6	S	9-F	0.0250	−10.79
7	S	CCH	0.0296	−10.69
8	S	9-Cl	0.0340	−10.60
9	S	OCH_3	0.0340	−10.60
10	S	H	0.0440	−10.44
11	S	I	0.0474	−10.39
12	O	Br	0.0473	−10.39
13	S	CN	0.0563	−10.29
14	O	I	0.0880	−10.01
15	S	CHO	0.1880	−9.54
16	O	CCH	0.4376	−9.02
17	S	9-CF_3	0.4850	−8.96
18	O	CH_3	0.9890	−8.52
19	O	CN	1.1396	−8.43
20	O	H	3.1550	−7.81
21	O	9-CF_3	5.9190	−7.42

[a] Y is attached to C8 unless otherwise indicated. [b] References 25 and 33. [c] Calculated from $\Delta G_{binding} = RT \ln IC_{50}$ at 310 K.

Binding Free Energy Calculations. Relative binding free energies were computed via thermodynamic cycles by performing TI simulations for pairs of ligands both in solvent and while bound to the non-nucleoside reverse transcriptase inhibitor (NNRTI) binding pocket in HIV-1 RT. For the solvation simulations, the hybrid molecule was solvated in a 20 Å cubic box of TIP3P[34] water molecules, and periodic boundary conditions were employed. For the bound simulations, the pdb structure, 1TVR,[35] was truncated so that only residues within ~20 Å of the crystallographic TIBO compound were retained, and the truncated protein−ligand system was solvated in a 37 Å sphere of water. Stochastic boundary conditions using a solvent boundary potential[36] of 25 Å with a 5 Å buffer region were employed; 244 and 6101 water molecules were explicitly included in the solvated and bound simulations, respectively. A nonbonded cutoff of 15 Å was used, and van der Waals switching and electrostatic force shifting functions were implemented between 10 and 12 Å. In all simulations, the temperature was maintained near 310 K by coupling the water molecules to a Langevin heat bath using a frictional coefficient of 62 ps^{-1}. Hydrogen bonds were restrained using the SHAKE[37] algorithm, and the time step was 2 fs. Heating phases were 10 ps regardless of the environment, while equilibration phases were 30 and 60 ps for the solvated and bound simulations respectively. The production runs were 300 ps, and the coordinates were saved every 300 steps. Simulations were performed for 11 different λ values: 0.025, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, and 0.975. Linear scaling by λ applied to all energy terms

except the bond and angle terms, which were treated at full strength regardless of λ value. All simulations were performed in triplicate, and the resulting mean and standard deviations are reported. All calculations were performed using the BLOCK module in the CHARMM molecular dynamics package v35a1 on dual 2.66 GHz Intel Quad Core Xeon processors. On a single processor, each solvated and bound simulation required 1 and 22 CPU hours, respectively. Because each simulation window was generated independently from the others, all simulations could be performed simultaneously on a computer cluster.

TIBO Parameter Assignments. Atom types for the TIBO compounds were assigned using MATCH (unpublished, D. L. Price and C. L. Brooks, III) with the extended CHARMM22[38] force field and CGenFF (private communication, K. Vanommeslaeghe and A. D. Mackerell). Where possible, bonded parameters that were absent in CGenFF and the CHARMM22 force field were approximated by those from the OPLS-AA force field taken from BOSSv4.2.[16] Bonded parameters for which there were no analogous assignments in existing CHARMM of BOSS parameter files were obtained by fitting ab initio energy calculations from Gaussian 03.[39] Equilibrium bond lengths, angles, and dihedrals were determined by energy minimization of the corresponding molecular fragments at the MP2 level of theory using the 6-31G* basis set. The respective force constants were determined by systematically distorting the structures away from the optimal values at the MP2 level of theory. van der Waals energy parameters (i.e., atomic radii, r_i, and energy well-depths, ε_i) were taken from analogous atom types in CHARMM22 and CGenFF.

Initial TIBO Charge Assignments. Initial partial charges were assigned using MATCH with the extended CHARMM22[38] force field and CGenFF. Partial charges for most of the Y-site fragments (i.e., hydrogen, alkyl, halides, aldehydes, and ethers) were adopted from their corresponding benzene derivatives. Partial charges for the nitrile and trifluorinated methyl fragments were adopted from alkylated derivatives. No optimized partial charges existed for the allyl and thioether fragments, so they were estimated from CHARMM parameters for alkene and methoxybenzene derivatives, respectively. CHARMM22 did not have a template that corresponded to the C═S fragment, so pairs of molecules that differed only at the X site (i.e., X = O or S) were geometry-optimized in Gaussian 03 at the MP2/6-311+G** level of theory, and partial charges were fit to the electrostatic potential using the CHELPG algorithm (Breneman and Wiberg, 1990). The largest differences in the CHELPG assigned partial charges (i.e., $\Delta q_{O\rightarrow S} > 0.2e$) between pairs of compounds were localized in five atoms in the five-membered ring (H−N−C(═O/S)−N); therefore, the partial charges of these five atoms in the X = O and X = S TIBO derivatives were approximated directly from the CHELPG charges. The charge assignments for the TIBO core are illustrated in Figure 2.

Alternate Fragment Charge Assignments. *CHELPG.* Partial charges for four Y fragments (i.e., Y = CN, CHO, OCH_3, and SCH_3) were reassigned on the basis of CHELPG assigned partial charges for MP2/6-311+G** geometry-optimized structures of the corresponding benzene derivatives in Gaussian 03. In "CHELPG2" simulations, the charges of the Y fragment atoms along with the charge of the ipso carbon atom were reassigned. In "CHELPG4" simulations, the charges of the ortho carbons and hydrogen atoms were also reassigned. The CHELPG partial charges for a given benzene derivative were uniformly offset such that the sum of the reassigned charges equaled zero. The reassigned charges are listed in Table 4 (note: the "2" denotes fragment+ipso site and the "4" denotes fragment+ipso site+2 ortho sites).

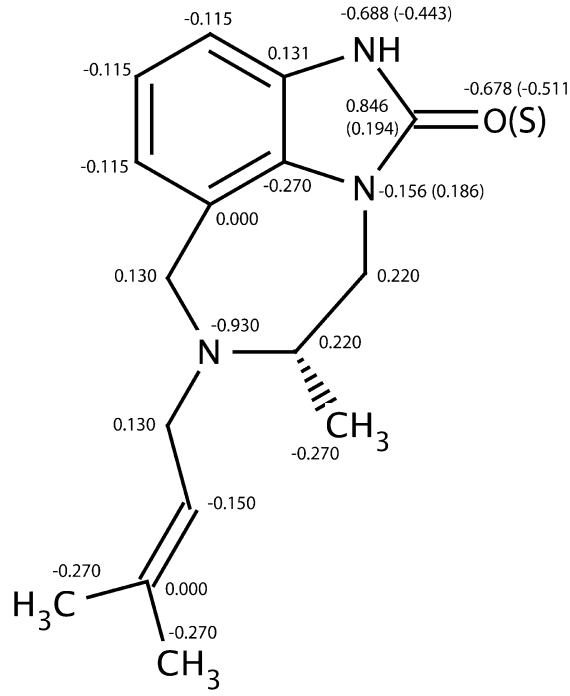


Figure 2. Initial partial charges assigned for heavy atoms on the TIBO core with X = O and Y = H. Partial charges associated with X = S are labeled in parentheses. Hydrogen atoms bonded to alkanes, alkenes, and aromatic carbons have partial charges of 0.09e, 0.150e, and 0.115e, respectively. The amide hydrogen has partial charges of 0.415e with X = O and 0.313e with X = S.

Alternate Fragment Charge Assignments. *CHopt.* Based on the strategy outlined by MacKerell et al.[38] for parametrizing ligands to be consistent with the CHARMM22 force field, the partial charges of the same four fragments at the Y site (i.e., Y = CN, CHO, OCH_3, and SCH_3) were optimized in the context of the corresponding benzene derivatives to yield molecular properties that were consistent with experimental hydration free energy data[40] as well as components of the dipole moment. The dipole moments were obtained in Gaussian 03 at the HF/6-31+G* level of theory and were scaled by 15%.

A Monte Carlo (MC) sampling strategy was employed in CHARMM in which many configurations of partial charges were evaluated for the atoms in the benzene derivative. In "CHopt2" simulations, the charges of the Y fragment atoms, along with the charge of the ipso carbon atom, were optimized. In "CHopt4" simulations, the charges of the ortho carbon and hydrogen atoms were also optimized. Partial charges of each of the atoms of interest were sampled such

that each $-1e < q_i < 1e$ (note: methoxy- and thiomethoxy-hydrogen atoms retained their charge of 0.09e throughout). Once trial partial charges were assigned, the components of the dipole were computed, and the atomic coordinates were minimized in vacuum for 100 steps using the Adopted Basis Newton−Raphson algorithm and then reminimized for 100 steps using the Steepest Descent algorithm using the GBMV implicit solvent model.[41,42] The hydration free energy was approximated as the difference between the solvent and vacuum energy minima.[20] The scoring function, S, for each configuration of partial charges, q, was defined by:

$$S_q = \frac{(\Delta\Delta G^q_{\mathrm{hydr}} - \Delta\Delta G^{\mathrm{target}}_{\mathrm{hydr}})^2}{2\sigma^2_{\Delta\Delta G}} + \frac{(\mu^q_x - \mu^{\mathrm{target}}_x)^2 + (\mu^q_y - \mu^{\mathrm{target}}_y)^2 + (\mu^q_z - \mu^{\mathrm{target}}_z)^2}{2\sigma^2_{\mu}}$$

where $\Delta\Delta G_{\mathrm{hydr}}$ denotes the hydration free energy relative to benzene, and μ_x, μ_y, and μ_z are the components of the molecular dipole ($\sigma_{\Delta\Delta G} = 0.25$ kcal/mol and $\sigma_\mu = 0.25$ D). The scoring function for CHopt4 optimizations included additional restraints to keep the partial charges near the initial MATCH- or CGenFF-assigned partial charges by imposing a fixed penalty of 5.5 whenever a trial partial charge deviated more than 0.1e from the initial charge. 40 000 trial configurations were sampled, and trial configurations were accepted with probability, P:

$$P = \min(1, \exp^{-(S_{\mathrm{trial}} - S_{\mathrm{previous}}/k_B T)})$$

The effective temperature, k_BT, was gradually decreased every n steps using an exponential cooling schedule, such that $k_BT_{t+1} = \alpha k_BT_t$. The initial temperatures and cooling schedules were optimized to ensure that the best-scored solutions were not dependent on the initial charge assignments (i.e., CHopt2, $n = 500$, $k_BT_0 = 200$, and $\alpha = 0.75$; CHopt4, $n = 1000$, $k_BT_0 = 20$, and $\alpha = 0.9$). The charge distributions that yielded the lowest-scored solutions were identified as the CHopt2 and CHopt4 charge models. The charge distribution that yielded the most poorly scored solution sampled was used as the "control" charge distribution.

Results and Discussion

Overall High Quality of Computed Binding Affinities. Relative binding affinities were computed for 44 pairs of TIBO compounds (Table 2). This data set encompasses 21 unique TIBO molecules and includes all transformations from Y = H and Y = CH_3. Eleven additional pairs were assessed: seven pairs that involved Y = halide→halide transformations and four pairs that involved X = O→S transformations. All possible combinations of pairwise relative binding affinities among the 21 TIBO compounds could theoretically be reconstructed from these representative calculations.

The average unsigned error (AUE) for the entire data set is 1.29 kcal/mol, and one-half of the TIBO pairs have computed binding affinities with individual errors of less than 1 kcal/mol, while the maximum unsigned error (MUE) is 5.38 kcal/mol. Figure 3 illustrates that by ranking the TIBO pairs by their difference in the predicted relative to the experimental binding free energy, the cumulative AUE for the top 89% of the data set is less than 1 kcal/mol. Computed binding affinities in this study have uncertainties on the order of ∼0.7 kcal/mol (Table 2). The majority of the simulations of the solvated ligands show standard deviations of less than 0.2 kcal/mol; most of the bound simulations show increased diversity yet with standard deviations of less than 0.6 kcal/mol. On the basis of a comparison between computed and experimental hydration free energies of small molecules, Mobley et al.[22] suggest that it will be difficult with current force fields to achieve average errors in binding affinities of less than 1 kcal/mol. Furthermore, it is estimated that experimental binding affinities have uncertainties of ∼0.5 kcal/mol[3]. Therefore, a cumulative AUE in our study of 1 kcal/mol represents quite a conservative threshold of "success".

Reproducibility of Binding Free Energies Is Size- and Class-Dependent. Large differences in the substituent size between pairs of compounds provide a more challenging context for adequately sampling relevant protein conformations in free energy calculations. Indeed, in this study, both the precision and the accuracy are deleteriously affected for simulations that involve large size differentials in the TIBO derivatives. Among most of the 11 pairs of TIBO derivatives whose transformations are less conservative in size (i.e., where the transformation at the Y-site involves an addition of more than one heavy atom or a transformation from H→Br or H→I), their standard deviations for simulations of the bound "arm" of the thermodynamic cycle are significantly larger than those for more conservative transformations. The AUE for simulations modeling larger size differentials is 2.3 kcal/mol, whereas the AUE for simulations involving more conservative transformations at the Y-site is 1.0 kcal/mol. These systematic errors are likely related to the λ-scaling scheme that was used in the TI calculations.[43] For example, simulations involving large substituent size differentials, the values of the integrand, $\langle \partial H/\partial\lambda \rangle$, and their standard deviations from independent trajectories at very low and very high λ values (i.e., $\lambda = 0.025$ and 0.975) are significantly larger than for those simulations involving substituents that are more similar in size. For structure-based drug design applications, longer simulation trajectories and more extensive soft-core scaling techniques[44] may be required to achieve adequate sampling with the anticipation of improving the quality of the estimated binding affinities.

Representatives from all classes of functional group at the Y-site, except for the thioether moiety, reliably reproduce experimental binding affinities for the TIBO NNRTIs. Figure 4 depicts the free energy errors for chemical transformations by functional group at the Y-site. A positive error indicates that the hydrogen or methyl TIBO derivative in the transformation (or the smaller halide in the case of Y = halide→halide transformations or oxygen in the case of X = O→S transformations) is overfavored relative to experiment. The eight pairs of compounds that contain only hydrogen, alkyl, and allyl groups at the Y site have individual errors that are less than 1.8 kcal/mol and have an AUE of

Table 2. Quality of Computed Binding Free Energies[a] Using Thermodynamic Integration for a Selection of Pairs of TIBO Compounds

				$\Delta\Delta G_{solv}$		$\Delta\Delta G_{bound}$			
pair no.	ligand 1 X, Y	ligand 2 X, Y	$\Delta\Delta G_{expt}$	mean	std dev	mean	std dev	$\Delta\Delta G_{calc}$	error
1	O, H	O, CH_3	−0.71	0.18	0.12	0.73	1.21	0.55	1.26
2	S, H	S, CH_3	−0.72	0.05	0.17	1.10	0.41	1.05	1.77
3	O, H	S, H	−2.63	3.75	0.17	1.92	0.43	−1.83	0.80
4	O, CH_3	S, CH_3	−2.64	3.98	0.07	1.67	0.50	−2.31	0.33
5	O, H	O, CCH	−1.21	14.44	0.22	13.10	0.26	−1.34	−0.13
6	S, H	S, CCH	−0.25	14.66	0.14	13.93	0.60	−0.73	−0.48
7	O, CH3	O, CCH	−0.50	14.63	0.08	14.09	1.10	−0.54	−0.04
8	S, CH_3	S, CCH	0.47	15.03	0.10	15.33	0.62	0.30	−0.17
9	S, H	S, F	−1.25	3.30	0.07	3.29	0.37	−0.01	1.24
10	S, 9H	S, 9F	−0.35	−9.31	0.07	−9.33	0.68	−0.03	0.32
11	S, CH_3	S, F	−0.53	2.50	0.05	0.00	0.27	−2.50	−1.97
12	S, F	S, Cl	−0.40	−0.88	0.08	0.34	0.26	1.23	1.63
13	S, 9F	S, 9Cl	0.19	2.44	0.02	2.56	0.26	0.12	−0.07
14	S, H	S, Cl	−1.43	2.04	0.05	3.16	0.24	1.12	2.55
15	S, 9H	S, 9Cl	−0.16	−7.35	0.02	−6.90	0.44	0.45	0.61
16	S, CH_3	S, Cl	−0.71	1.47	0.03	0.52	0.13	−0.95	−0.24
17	S, Cl	S, Br	−0.22	−0.08	0.02	0.79	0.31	0.87	1.09
18	O, H	O, Br	−2.58	3.55	0.11	4.63	0.60	1.08	3.66
19	S, H	S, Br	−1.65	1.62	0.19	2.53	1.01	0.91	2.56
20	O, CH_3	O, Br	−1.87	3.20	0.08	2.99	0.14	−0.21	1.66
21	S, CH_3	S, Br	−0.93	1.23	0.04	1.65	0.52	0.42	1.35
22	O, Br	S, Br	−1.70	3.77	0.02	1.90	0.21	−1.87	−0.17
23	O, Br	O, I	0.38	−0.76	0.03	0.05	0.05	0.81	0.43
24	S, Br	S, I	1.70	−0.50	0.08	−0.21	0.31	0.29	−1.41
25	O, H	O, I	−2.20	2.49	0.09	4.64	0.88	2.15	4.35
26	S, H	S, I	0.05	0.53	0.10	3.14	0.46	2.61	2.56
27	O, CH_3	O, I	−1.49	2.33	0.01	2.97	0.47	0.64	2.13
28	S, CH_3	S, I	0.77	0.63	0.16	1.76	0.33	1.13	0.36
29	O, I	S, I	−0.38	3.52	0.12	2.40	0.30	−1.11	−0.73
30	O, 9H	O, 9CF_3	0.39	2.68	0.19	4.10	0.19	1.41	1.02
31	S, 9H	S, 9CF_3	1.48	2.65	0.17	4.64	0.13	1.99	0.51
32	S, 9F	S, 9CF_3	1.83	12.95	0.11	14.62	0.09	1.68	−0.15
33	S, 9Cl	S, 9CF_3	1.64	10.69	0.04	12.70	0.33	2.01	0.37
34	O, H	O, CN	−0.62	8.42	0.20	9.30	0.81	0.88	1.50
35	S, H	S, CN	0.15	8.62	0.08	10.57	1.34	1.95	1.80
36	O, CH_3	O, CN	0.09	8.68	0.05	8.49	0.52	−0.19	−0.28
37	S, CH_3	S, CN	0.87	9.16	0.09	9.10	0.31	−0.06	−0.93
38	O, CN	S, CN	−1.86	4.00	0.18	2.17	0.21	−1.83	0.03
39	S, H	S, CHO	0.90	7.23	0.06	8.38	0.87	1.15	0.25
40	S, CH_3	S, CHO	1.62	7.27	0.11	6.73	0.34	−0.54	−2.16
41	S, H	S, OCH_3	−0.16	9.03	0.08	11.53	0.97	2.50	2.66
42	S, CH_3	S, OCH_3	0.54	9.31	0.18	10.33	0.87	1.02	0.48
43	S, H	S, SCH_3	−1.34	2.48	0.04	6.53	0.29	4.04	5.38
44	S, CH_3	S, SCH_3	−0.62	3.10	0.12	5.52	0.35	2.43	3.05

[a] Means and standard deviations are reported for three independent sets of simulations where each set includes 11 simulations at λ values: 0.025, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, and 0.975.

0.6 kcal/mol. The 13 pairs of compounds that contain either fluoride, chloride, or trifluorinated methyl groups at the Y position demonstrate random errors in the computed binding free energies and collectively have an AUE of 0.9 kcal/mol. The quality of the modeled bromide and iodide substituents at the Y-site is degraded relative to the rest of the data set, but this is primarily due to the large size differential in four of the simulations. The AUE of the latter four simulations that involve transformations from hydrogen to either bromide or iodide at the Y-site is 3.3 kcal/mol, whereas the AUE for all other bromide and iodide transformations is 1.0 kcal/mol. The Y = H→OCH_3 transformation also suffers from a relatively large error of 2.7 kcal/mol, but the Y = CH_3→OCH_3 transformation is accurately computed with an error of 0.5 kcal/mol. The majority of the binding affinities among the nitriles and aldehydes TIBO derivatives are computed reliably with AUEs of 0.9 and 1.2 kcal/mol, respectively. By contrast, simulations of pairs of compounds that contain the thioether fragment yield the largest individual and collective errors in the data set. Specifically, the Y = H→SCH_3 and Y = CH_3→SCH_3 transformations systematically underestimate the relative binding affinities of the thioether TIBO derivative and have errors of 5.4 and 3.1 kcal/mol, respectively.

The success of the hydrogen and alkyl TIBO derivatives is not surprising given that these atoms have analogues in well-parametrized amino acid side chains in the CHARMM22 force field. The high quality of the binding affinities for the halide and methoxy TIBO derivatives validates the bonded and nonbonded parameters that were recently optimized by Vanommeslaeghe and Mackerell for methoxybenzene and the halobenzenes. The consistently large and systematic errors for the thioether derivative are not surprising given that, in the absence of parameters for thiomethoxybenzene,

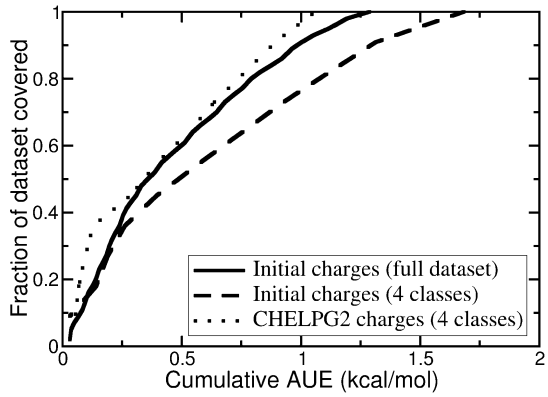


Figure 3. Cumulative errors in computed binding affinities for pairs of TIBO compounds ordered by absolute error. The full data set includes all 44 pairs of TIBO compounds listed in Table 2. The "4 classes" includes 11 pairs of TIBO compounds (pairs 34−44 in Table 2), which involve transformations for Y = CN, CHO, OCH_3, or SCH_3.

their initial charges were estimated from methoxybenzene. This finding suggests that optimization efforts could be targeted toward improving the parametrization of the thioether TIBO derivatives.

Overall, the success of the X = O→S transformations indicates that the balance of the charge distributions between the oxygen and sulfur TIBO derivatives is reasonable. Yet, the large percentage of positive errors for transformations at the Y-site (29 out of 41 cases) suggests a systematic overfavoring of the hydrogen or methyl substituent (or the smaller halide in the Y = halide→halide transformations) relative to experiment. These predominantly represent the favoring of the smaller of the two substituents under consideration in a given simulation. This bias could be due to charges associated with the amide hydrogen that results in a strong hydrogen bond with the K101 backbone carbonyl oxygen at the mouth of the binding pocket. The strength of this hydrogen bond may prevent sufficient relaxation of the TIBO compound such that the interactions with the protein environment at the other end of the binding pocket are too restrictive and thus unfavorable for the bulkier substituent.

Charge Optimization Strategies Improve Thioether Computed Binding Affinities. Based on these results, the atomic partial charges associated with the thioether fragment were targeted for further optimization. Partial charges of the nitrile, aldehyde, and ether fragments were also optimized as controls to confirm the transferability of any proposed charging scheme across a variety of functional groups. To ensure that these charge distributions would be generalizable beyond the TIBO compounds, each of these four functional groups was investigated as a substituent at a single site on a benzene ring.

The four optimization strategies that have been investigated explore how the charges are assigned and the extent to which a given fragment influences that charge distribution in the rest of the molecule. In the first strategy (CHELPG), charges were adopted from the CHELPG charges that were fit to the electrostatic potential. In the second strategy (CHARMM optimization − CHopt), partial atomic charges were optimized via a Monte Carlo procedure to yield good agreement with the components of the QM molecular dipole as well as experimental hydration free energies relative to benzene. As a first approximation (CHELPG2 and CHopt2), partial charge distributions are assumed to be local in nature, and thus charge assignments are limited to atoms in the functional group and the ipso carbon on the benzene ring. More extensive charge delocalization was also investigated (CHELPG4 and CHopt4) such that the charge assignments for each of these four functional groups were specific for the ortho carbon and hydrogen atoms on the benzene ring as well as the atoms in the functional group and the ipso carbon (note: the "2" denotes functional group+ipso site and the "4" denotes functional group+ipso site+2 ortho sites). From the CHopt2 MC trajectories, a "control" charging scheme was identified, which yielded the poorest fit to the targeted physical properties of the model benzene derivatives. Table 3 describes the molecular properties that result from these different charge distributions in the respective benzene derivatives. Table 4 summarizes the errors in the relative binding affinities that were recomputed for these four classes of TIBO derivatives.

The CHELPG2 charge distribution for thioether benzene is similar to that of the initial charges; yet, it yields a better estimate of the hydration free energy than the initial charges. When these CHELPG2 charges are transferred to the TIBO compound, there is a marked improvement in the thioether computed binding free energies; the error for each of the two thioether transformations improves by at least 1 kcal/mol when the initial charge estimates are replaced by the CHELPG2 charges. The CHopt2 charge model also has an improved fit to the experimental hydration free energy and QM dipole moment relative to the initial charge model. This charge distribution in the TIBO derivative elicits an improvement in the computed binding affinities by 0.5−2.1 kcal/mol relative to the initial charge, although it is not overall as favorable as the result for the CHELPG2 charge model. Increasing the scope of the charge delocalization in the CHELPG4 and CHopt4 models yields better agreement with the targeted molecular properties for the thioether benzene derivatives; yet, these charge models do not improve the computed binding affinities for the corresponding TIBO derivatives (errors of 2.7 and 4.1 kcal/mol) over the CHELPG2 and CHopt2 models. However, the quality of the CHopt4 charge model may be unduly hindered because the initial charges to which the CHopt4 partial charges are restrained were approximated from the methoxybenzene charge distribution. The "control" charge model, which has the poorest agreement with the targeted molecular properties of any of the charge models, exhibits the worst binding free energies when it is transferred to the TIBO compound (errors of 4.2 and 5.4 kcal/mol).

Charge Optimization Strategies Adequate for Nitriles, Aldehydes, and Ethers. Results from the other three classes of fragments (Y = CN, CHO, and OCH_3) demonstrate that charge models obtained from schemes either that fit charges to the electrostatic potential or that optimize charge distributions to mimic hydration free energies and molecular dipole moments are sufficient to compute reliable estimates of binding free energies. The AUEs for each of

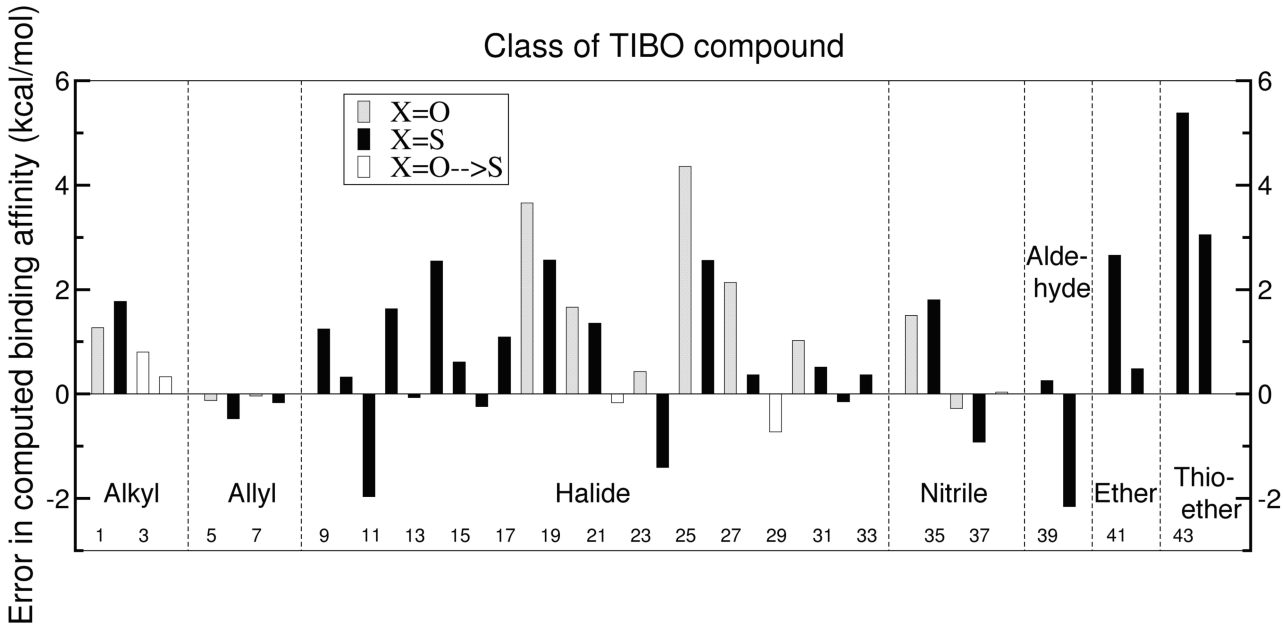


Figure 4. Errors in binding affinities enumerated by pair number (see Table 2) and categorized by identity of the Y fragment.

these four functional groups and charge models range from 0.3 to 2.0 kcal/mol. The CHELPG2 models perform favorably in which all but one computed binding free energy has an error of less than 1.4 kcal/mol. Figure 3 illustrates the significant improvement in the overall quality of the computed binding affinities for the CHELPG2 models relative to the initial charges for the functional groups investigated. The overall AUE for these 9 pairs of TIBO compounds improves from 1.68 to 1.06 kcal/mol, and the MUE is reduced from 2.66 to 1.80 kcal/mol.

Even though CHELPG2 and CHELPG4 charges were assigned from the same set of CHELPG charges that were fit to the electrostatic potential of the model benzene compounds, the atomic charges differ slightly among the CHELPG2 and CHELPG4 charge sets due to approximations that were introduced to restrain the sum of the reassigned charges to be zero. For each substituent, the CHELPG4 charges are within 0.06e of the CHELPG2 assigned charges for the fragment and the ipso carbon, although most vary by less than 0.02e. It was anticipated that these more delocalized charge distributions in the CHELPG4 models, which increase the specificity of the context of the functional group in the benzene derivatives relative to the CHELPG2 models, would improve the quality of the corresponding TIBO binding affinities. In fact, relative to the CHELPG2 models, CHELPG4 charge distributions tend to degrade the quality of the computed binding affinities for the nitriles, aldehydes, and ethers derivatives.

The alternative CHopt2 scheme for assigning localized partial charges results in charge distributions that differ substantially from the initial as well as the CHELPG2 charges; these resulting charge distributions yield relative hydration free energies and components of the QM dipole moments that are closer to the target values than either the initial or the CHELPG2 charges. We assumed that the set of charges that are optimized by this scheme would result in higher quality computed binding affinities for the respective TIBO compounds due to the increased ability of the charges to mimic physical properties of the corresponding model benzene compound. Indeed, these CHopt2 charge distributions slightly improve the binding affinities for most of the pairs of TIBO derivatives relative to the initial charge models. The one exception is the relatively large error for the Y = H→CHO transformation. Interestingly, the partial charge assignments for Y = CN and Y = OCH_3 differ by up to 0.3e relative to the initial charges; yet the high quality of the binding affinities is still achieved.

By permitting optimization of the charges of the carbon and hydrogen atoms in the ortho position, the CHopt4 partial charge assignments in the benzene derivatives yielded better agreement with the experimental hydration free energies and QM dipole moments than the CHopt2 charges, although restraining the partial charges to the initial charges resulted in poorer fits with the targeted molecular properties for Y = CN. With the increased ability of the charge distributions of the TIBO fragments to mimic critical molecular properties, it was assumed that the CHopt4 charges would result in greater improvements in the corresponding binding affinities than the other models. In fact, the quality of the computed binding affinities tend to be degraded relative to the CHopt2 charge models, although both transformations involving Y = CHO achieve remarkably low errors (AUE of 0.3 kcal/mol). The AUE is degraded slightly from 0.6 to 1.3 kcal/mol and from 1.2 to 1.8 kcal/mol for Y = CN and Y = OCH_3, respectively.

Given the relative success of the charge optimization schemes for these nitrile, aldehyde, and ether functional groups and the inability of these schemes to improve the quality of the binding affinities for the thioether TIBO derivatives beyond 2 kcal/mol, we suggest that further optimization of the other nonbonded parameters (i.e., atomic radius and energy well-depth) is likely required in conjunction with improved partial charges, but is beyond the scope of this study.

Table 3. Partial Charge Assignments from Various Charge Optimization Schemes and Their Respective Molecular Properties[a]

			charging schemes					
benzene substituent		atom name/targets	initial	CHELPG2	CHopt2	CHELPG4	CHopt4	control
Y = SCH_3	charges:	CA	0.220	0.206	0.039	0.218	0.291	0.468
		S	−0.390	−0.305	−0.186	−0.292	−0.292	−0.432
		CT3	−0.100	−0.005	−0.122	0.007	−0.110	−0.305
		HA	0.090	0.035	0.090	0.046	0.090	0.090
		CA	−0.115	−0.115	−0.115	−0.185	−0.214	−0.115
		HP	0.115	0.115	0.115	0.150	0.134	−0.115
	$\Delta\Delta G_{hydr}$	−1.83 (expt)	−3.27	−1.20	−1.37	−1.86	−1.80	−5.46
	μ_x	−0.04 (HF)	−1.42	−1.76	−0.03	−0.65	−0.11	−4.29
	μ_y	1.72 (HF)	2.34	1.52	1.74	1.70	1.93	1.35
Y = CN	charges:	CA	0.130	0.097	0.425	0.084	0.132	−0.475
		CN	0.400	0.370	−0.238	0.357	0.308	−0.451
		NC	−0.530	−0.467	−0.187	−0.480	−0.440	0.024
		CA	−0.115	−0.115	−0.115	−0.124	−0.019	−0.115
		HP	0.115	0.115	0.115	0.143	0.308	0.115
	$\Delta\Delta G_{hydr}$	−2.66 (expt)	−5.30	−3.70	−3.04	−4.47	−3.70	−5.53
	μ_x	−5.84 (HF)	−4.49	−3.90	−4.58	−4.13	−2.80	2.82
Y = CHO	charges:	CA	0.120	0.051	0.252	0.055	0.177	−0.264
		CD	0.160	0.465	−0.017	0.469	0.099	0.581
		O	−0.330	−0.536	−0.332	−0.531	−0.378	−0.530
		HR1	0.050	0.021	0.098	0.025	0.092	0.213
		CA	−0.115	−0.115	−0.115	−0.152	−0.016	−0.115
		HP	0.115	0.115	0.115	0.142	0.022	0.115
	$\Delta\Delta G_{hydr}$	−3.18 (expt)	−1.98	−5.63	−3.31	−5.83	−3.30	−11.39
	μ_x	3.19 (HF)	2.28	3.16	2.68	3.09	2.92	3.53
	μ_y	−3.05 (HF)	−2.02	−2.11	−2.80	−1.78	−3.01	0.71
Y = OCH_3	charges:	CA	0.220	0.458	−0.022	0.513	0.318	0.020
		O	−0.390	−0.515	−0.093	−0.460	−0.290	−0.503
		CT3	−0.100	0.148	−0.155	0.203	−0.065	0.213
		HA	0.090	−0.030[b]	0.090	0.025	0.090	0.09
		CA	−0.115	−0.115	−0.115	−0.326	−0.135	−0.115
		HP	0.115	0.115	0.115	0.160	0.019	0.115
	$\Delta\Delta G_{hydr}$	−0.20 (expt)	−1.92	−2.43	−0.57	−2.17	−0.43	−7.73
	μ_x	0.68 (HF)	−0.75	−3.26	0.49	1.07	0.53	2.02
	μ_y	1.53 (HF)	1.64	1.15	0.93	1.57	1.48	2.71

[a] Experimental hydration free energies ($\Delta\Delta G_{hydr}$) relative to benzene in kcal/mol taken from ref 40. Computed relative hydration free energies are approximated from the difference between GBMV and vacuum energy-minimized energies. Dipole moments are reported in units of Debye from the standard Gaussian 03 orientation, and their HF/6-31+G* values have been scaled by 15%. [b] The negative charges assigned to the methyoxy hydrogen atoms are a result of the offset factor used to require that the overall charge of the reassigned atoms sums to zero.

Table 4. Effect of the Partial Charge Distributions on the Quality of the Computed Relative Binding Free Energies[a]

		errors in charging schemes					
ligand 1 X, Y	ligand 2 X, Y	initial	CHELPG2	CHopt2	CHELPG4	CHopt4	control
O, H	O, CN	1.50	−0.23	0.03	1.73	1.63	4.18
O, CH_3	O, CN	1.80	−0.08	0.91	1.95	1.01	3.16
S, H	S, CN	−0.28	0.05	−0.73	0.33	0.18	3.44
S, CH_3	S, CN	−0.93	−0.97	−0.96	−0.42	−0.59	2.08
O, CN	S, CN	0.03	−0.15	−0.28	0.42	−2.95	−2.70
	AUE(Y = CN):	0.9	0.3	0.6	1.0	1.3	3.1
S, H	S, CHO	0.25	1.40	1.87	3.04	0.39	4.50
S, CH_3	S, CHO	−2.16	−1.30	−2.05	−0.14	−0.19	1.23
	AUE(Y = CHO):	1.2	1.4	2.0	1.6	0.3	2.9
S, H	S, OCH_3	2.66	1.79	2.15	2.14	3.31	3.83
S, CH_3	S, OCH_3	0.48	0.92	−0.15	0.39	0.28	3.09
	AUE(Y = OCH_3):	1.6	1.4	1.2	1.3	1.8	3.5
S, H	S, SCH_3	5.38	2.72	3.26	4.05	3.91	5.39
S, CH_3	S, SCH_3	3.05	2.04	2.56	2.73	2.94	4.22
	AUE(Y = SCH_3):	4.2	2.4	2.9	3.4	3.4	4.8

[a] Errors in the computed relative binding free energies are reported in kcal/mol ($\Delta\Delta G_{calc} - \Delta\Delta G_{expt}$).

The Importance of Physically Meaningful Charge Distributions. A "control" charging scheme was selected for each functional group to ascertain the importance of physically relevant charge distributions for effectively modeling binding affinities for these TIBO derivatives. The "control" charging schemes exhibit poor agreement with experimental hydration free energies and QM dipole moments and yield very poor quality results among the TIBO

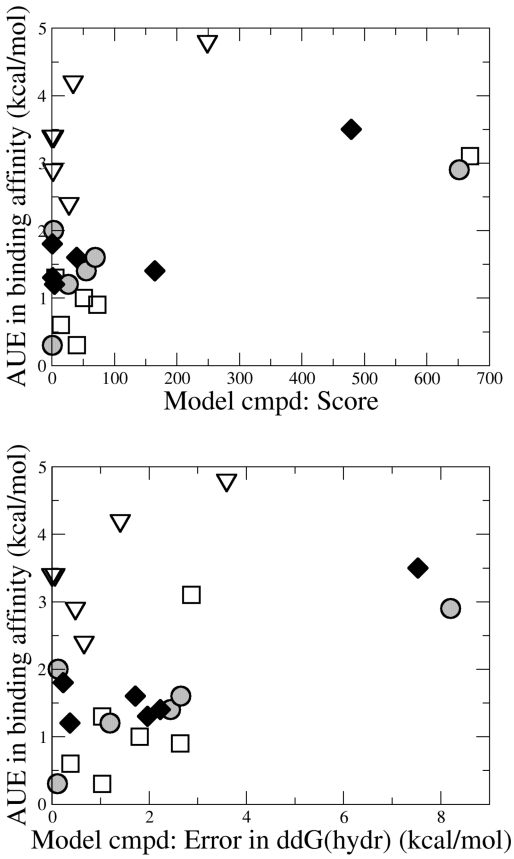


Figure 5. Correlation between the physical properties of the model benzene derivatives and the quality of the resulting binding free energies of the corresponding TIBO compounds for Y = CN (open squares), Y = CHO (shaded circles), Y = OCH_3 (filled diamonds), and Y = SCH_3 (open triangles). The score (top) that is used in the MC simulations can be understood as a "penalty" such that the lower is the number, the better is the agreement between the targeted molecular properties. The error in the hydration free energy (bottom) is approximated from the difference between GBMV and vacuum-minimized energies.

binding affinities. In fact, the AUE for each class degrades to more than 2.9 kcal/mol, and all but one individual binding free energy has an error of more than 2 kcal/mol. Figure 5 illustrates the correlation that is observed between the quality of the charge distributions in the modeled compound (as measured by the score used in the MC optimization or by the error in the hydration free energy relative to benzene) and the quality of the computed binding affinities in the corresponding TIBO derivatives. From Figure 5, it is clear that reasonable physical properties for model compounds (i.e., low MC scores or low errors in hydration free energies) are required for achieving high-quality binding free energies. In fact, the lowest errors in predicted binding affinities result from charge distributions that predict hydration free energies of their model compounds within 1 kcal/mol. However, achieving accurate experimental hydration free energies or good scores in model compounds does not guarantee success in reproducing experimentally binding affinities as is demonstrated by the consistently poor performance of any of the charge distributions for the thiomethoxy TIBO derivatives. A larger set of data for each functional group as well as additional classes of compounds would need to be explored to more fully describe the relationship between molecular properties and the quality of the computed binding affinities in the regime where the charge distributions are physically meaningful (i.e., the distributions spanned by the CHELPG and CHopt models in this study). In addition, other charging schemes like the scaled CM1A charge model and the AM1-BCC charges, which have proved effective for modeling with OPLS-AA and AMBER, respectively, could also be investigated for their compatibility with the CHARMM force field and are under investigation in our group. It is also worth emphasizing that the inclusion of experimental hydration free energies and scaled dipole moments in our scoring function was designed to be compatible with other CHARMM22 force field parameter development efforts,[38] so likely minor modifications or inclusion of other key molecular properties would be required for the appropriate transferability of measures of model quality to other biomolecular force fields.

Developing Charge Distributions in Model Compounds. In molecular modeling, there is always the need to balance chemical rigor with computational efficiency. This is especially relevant in a discussion about ligand parametrization where the need for transferability of parameters to novel drug-like molecules must be held in tension with the demand for high-quality estimates of binding affinities in investigating these new compounds in silico. At one end of the spectrum, new parameters could be optimized for each novel compound under investigation. While this strategy may produce more reliable results, given the enormity of chemical space, it is too time-consuming to be realistically pursued. Recently, Åqvist and co-workers investigated the plausibility of adopting charge distributions for complete drug molecules that were estimated from automated semiempirical and ab initio methods. Their study showed that several charge schemes (including CHELPG charge distributions) were reasonably compatible with the OPLS-AA force field for computing binding affinities with linear interaction energy (LIE) models.[45] While these charge schemes did not achieve the same level of success as OPLS-AA-optimized charges, they suggest that these automated schemes could be used as reasonable approximations in high-throughput calculations.

An alternative strategy involves optimizing charge distributions on fragments or model compounds that could be used to build up any new molecule. Maciel and Garcia have examined how the molecular context affects charge assignments to identify the smallest context that is required to reliably reproduce CHELPG charge assignments from a molecule's constitutive fragments.[46] Using a large test set of 324 molecules, they determined that five or more heavy-atom neighbors are typically required for accurately transferring charge assignments from one molecule or molecular fragment to another. This "five-atom neighborhood" could represent the ideal conditions for partial charge transferability; however, it is still significantly beyond the scope of current fragment library development efforts.

Generally, automated parameter assignment schemes use atom types that are obtained by matching molecular fragments that describe functional groups covered by the force field. A molecular fragment is associated with a given set of BCIs that

describe the magnitude and direction of the partial charges associated with a covalent bond between any two atom-types. The high quality of the individual and collective binding affinity results for the well-parametrized initial charges, as well as the CHELPG2 and CHopt2 schemes, is a promising indication that bond-charge increment rules and their associated partial charge distributions that are derived from physically meaningful charge distributions on model compounds can be transferred successfully to novel compounds. Furthermore, these results suggest that the important "neighborhood" is relatively local in nature. Of course, there may be exceptions to this finding, as demonstrated by the halide substituents whose influence is modeled to extend to the charges on the ortho carbon and hydrogen atoms in the recently optimized CGenFF parameters. However, for the most part, as fragment libraries are developed, the bond-charge increment rules that include the identity of the atoms that attach fragments to one another (e.g., the ipso carbon on the benzene ring) will likely be sufficient.

In our opinion, the CHopt2 charge models do not perform sufficiently well to warrant the added computational expense that is required to develop these charge distributions. Because of the success of the CHELPG2 charge distributions in computing binding free energies and how readily the model may be obtained (within minutes to a few hours on a standard desktop machine), this work supports the use of charge distributions that are derived from the ESP of model compounds for rapidly generating new bond-charge increments to investigate novel compounds or to expand current fragment libraries.

Conclusions

Here, we have performed a systematic assessment of the quality of binding affinities than can be achieved with current and recently optimized CGenFF parameters for a large series of non-nucleoside inhibitors that bind to HIV-1 RT. Thermodynamic integration simulations were performed to compute relative binding affinities for 44 pairs of TIBO compounds, which cover 21 unique molecules. These calculations achieve a high level of success with average errors in the binding affinities of 1.29 kcal/mol for the entire data set, and one-half of the pairs of compounds exhibit individual errors of less than 1 kcal/mol. While representatives of each of the CGenFF functional groups that were tested performed well, the quality of the results depended significantly on the size of the modeled substituents. TI simulations that modeled the transformation between substituents of similar sizes tended to be more successful (AUE of 1.0 kcal/mol for 33 pairs) than transformations that involved larger size differentials (AUE of 2.3 kcal/mol for 11 pairs). Binding affinities for TIBO derivatives containing alkyl, allyl, aldehydes, nitriles, trifluorinated methyl, and conservative halide transformations were reliably computed and had AUEs between 0.6 and 1.2 kcal/mol. By contrast, the thioethers whose partial charge assignments were approximated from methoxybenzene demonstrated large and systematic errors that consistently overfavored the binding of the hydrogen or methyl TIBO derivative relative to the thioether counterparts; their individual errors were greater than 3 kcal/mol, and the AUE was 4.2 kcal/mol.

Because of its large and systematic errors and the fact that thioether CHARMM parameters have not yet been developed, parameters of the thioether TIBO compound was targeted for optimization. Three additional classes of compounds were selected as controls: the nitriles, aldehydes, and ethers. We have investigated how different charging schemes for small molecules in conjunction with the CHARMM force field impact the quality of the computed binding affinities for this subset of TIBO compounds. The four charge distribution schemes that we tested each improved the quality of the computed binding affinities for the thioether TIBO derivative relative to its initial charges and performed reasonably well for the nitriles, aldehydes, and ethers. The CHELPG2 charge optimization scheme, which adopted localized partial charges that were fit to the QM electrostatic potential of model benzene, yielded the smallest average binding affinity error among the pairs of TIBO compounds investigated; the AUE of the 11 pairs of TIBO compounds was reduced from 1.7 kcal/mol with the initial charge distributions to 1.1 kcal/mol with the CHELPG2 charge assignments, and the MUE for these 11 pairs was reduced from 5.4 to 2.7 kcal/mol. By contrast, the "control" charge distributions that specifically did not mimic experimental or QM target molecular properties for the model benzene compounds resulted in extremely poor quality binding affinities with an AUE of 3.4 kcal/mol and MUE of 5.4 kcal/mol across the 11 pairs of TIBO compounds. Because the thioethers were still consistently underfavored relative to their alkyl counterparts in each of the charge optimization schemes, we suggest that other nonbonded parameters will likely need to be optimized before further improvements in the corresponding binding affinities are observed.

This study demonstrates the quality of recently developed CGenFF parameters as well as the advantage of using model compounds to derive physically meaningful charge distributions in the absence of parametrized bond-charge increments for a given compound. Because of the high quality of the binding affinities computed using the CHELPG2 partial charge assignments, we suggest that this kind of charge optimization strategy can be used either to rapidly generate charge distributions for specific drug-like models of interest or to expand bond-charge increments and fragment libraries of current force fields.

Acknowledgment. We thank Kenno Vanommeslaeghe and Alex D. Mackerell, Jr., for providing the CGenFF parameters for halobenzenes and methoxybenzenes. This research was supported by the National Institutes of Health (GM37554).

References

(1) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(2) Taft, C. A.; Da Silva, V. B.; Da Silva, C. H. *J. Pharm. Sci.* **2008**, *97*, 1089–1098.

(3) Foloppe, N.; Hubbard, R. *Curr. Med. Chem.* **2006**, *13*, 3583–3608.

(4) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(5) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.

(6) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(7) Mackerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(8) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730–748.

(9) Guvench, O.; MacKerell, A. D., Jr. *Methods Mol. Biol.* **2008**, *443*, 6388.

(10) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(11) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

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

(13) Schuttelkopf, A. W.; van Aalten, D. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–1363.

(14) van Aalten, D. M.; Bywater, R.; Findlay, J. B.; Hendlich, M.; Hooft, R. W.; Vriend, G. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 255–262.

(15) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res.* **2007**, *35*, 522–525.

(16) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(17) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *188*, 11225–11236.

(18) van Gunsteren, W. F.; Berendsen, H. J. C. *BIOMOS*; Groningen, The Netherlands, 1987.

(19) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(20) Price, D. J.; Brooks, C. L., III. *J. Comput. Chem.* **2005**, *26*, 1529–1541.

(21) Udier-Blagovic, M.; Morales De Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.

(22) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

(23) Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. *J. Comput. Chem.* **1993**, *14*, 1240–1249.

(24) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. *J. Am. Chem. Soc.* **2001**, *123*, 5221–5230.

(25) Smith, R. H., Jr.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A.; Michejda, C. J.; Kroeger Smith, M. B. *J. Med. Chem.* **1998**, *41*, 5272–5286.

(26) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. *J. Chem. Theory Comput.* **2007**, *3*, 256–277.

(27) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

(28) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6290–6295.

(29) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.

(30) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(31) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(32) Halgren, T. A. *J. Comput. Chem.* **1998**, *17*, 520–552.

(33) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; et al. *J. Med. Chem.* **1995**, *38*, 794–802.

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

(35) Das, K.; Ding, J.; Hsiou, Y.; Clark, A. D., Jr.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A.; Boyer, P. L.; Clark, P.; Smith, R. H., Jr.; Kroeger Smith, M. B.; Michejda, C. J.; Hughes, S. H.; Arnold, E. *J. Mol. Biol.* **1996**, *264*, 1085–1100.

(36) Brooks, C. L., III.; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24*, 843–865.

(37) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Phys.* **1977**, *34*, 1311–1327.

(38) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

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

(40) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Sol. Chem.* **1981**, *10*, 563–595.

(41) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(42) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(43) Pitera, J. W.; van Gunsteren, W. F. *Mol. Simul.* **2002**, *28*, 45–65.

(44) Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(45) Wallin, G.; Nervall, M.; Carlsson, J.; Aqvist, J. *J. Chem. Theory Comput.* **2009**, *5*, 380–395.

(46) Maciel, G. S.; Garcia, E. *Chem. Phys. Lett.* **2006**, *420*, 497–502.

CT900079T

JCTC Journal of Chemical Theory and Computation

An Adaptive Fast Multipole Boundary Element Method for Poisson−Boltzmann Electrostatics

Benzhuo Lu,*,† Xiaolin Cheng,‡ Jingfang Huang,§ and J. Andrew McCammon||

Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3250, Department of Chemistry & Biochemistry, Center for Theoretical Biological Physics, Department of Pharmacology, Howard Hughes Medical Institute, University of California, San Diego, California 92093

Received February 16, 2009

Abstract: The numerical solution of the Poisson−Boltzmann (PB) equation is a useful but a computationally demanding tool for studying electrostatic solvation effects in chemical and biomolecular systems. Recently, we have described a boundary integral equation-based PB solver accelerated by a new version of the fast multipole method (FMM). The overall algorithm shows an order *N* complexity in both the computational cost and memory usage. Here, we present an updated version of the solver by using an adaptive FMM for accelerating the convolution type matrix-vector multiplications. The adaptive algorithm, when compared to our previous nonadaptive one, not only significantly improves the performance of the overall memory usage but also remarkably speeds the calculation because of an improved load balancing between the local- and far-field calculations. We have also implemented a node-patch discretization scheme that leads to a reduction of unknowns by a factor of 2 relative to the constant element method without sacrificing accuracy. As a result of these improvements, the new solver makes the PB calculation truly feasible for large-scale biomolecular systems such as a 30S ribosome molecule even on a typical 2008 desktop computer.

1. Introduction

Electrostatic interactions play essential roles in many biological processes, such as enzymatic catalysis, molecular recognition, and bioregulation. Over the past three decades, the Poisson−Boltzmann (PB)-based continuum electrostatic calculation has become a common tool in theoretical studies of biomolecular systems such as proteins and DNAs in aqueous solutions. Many of these PB solvers rely on numerical solution of the PB equation. Among them, the PB solvers based on the finite difference methods, including DelPhi, GRASP, MEAD, UHBD, and the PBEQ,[1] have gained wide popularity, most likely due to their ease of implementation. A finite volume/multigrid PB solver APBS also enjoys increasing popularity over biochemistry and biophysical communities recently.[2,3] To our knowledge, APBS is the first program to enable distribution of PB calculations to a great number of processors, thus allowing extremely large-scale systems to be computed.

On the other hand, algorithms based on the boundary integral equation (BIE) approach have shown great promise for their efficiency on scaling and memory requirements.[4−6] These methods rely on Green's theorem and potential theory to recast the linear PB equation into a set of boundary integral equations that need to be solved only on the surface of the

* To whom correspondence should be addressed. E-mail: bzlu@lsec.cc.ac.cn.
† Chinese Academy of Sciences.
‡ Oak Ridge National Laboratory.
§ University of North Carolina at Chapel Hill.
|| University of California, San Diego.

10.1021/ct900083k CCC: $40.75
Published on Web 05/21/2009

molecule. Therefore, the number of unknowns is reduced relative to the volumetric discretization in finite difference and finite element methods. This surface integral equation idea is not new and was applied in the boundary element methods (BEM) in the early 1970s for different kinds of problems. Unfortunately most previous BEM implementations used Gauss eliminations to solve the resulting linear system. Even when a Krylov subspace based iterative solver was used for acceleration, the BEM approach was still limited by the cost associated with numerous surface integrations (matrix vector multiplications) that require an order $\sim N^2$ operations for a system with N surface elements. In the last 20 years or so, however, many fast algorithms have been introduced to efficiently evaluate these convolution type surface integrations, examples include the FFT-based algorithms (such as the precorrected FFT[7,8] and particle mesh Ewald methods[9,10]) and the multipole expansion-based techniques (such as the tree code[11,12] and fast multipole methods[13−17]). In particular, we want to mention our recent combination of the new version of the fast multipole method with the BEM formulation for PB equation, which has been shown numerically to be faster than existing PB solvers based on the finite-difference method for relatively large systems.[6]

However, our earlier implementation of the BEM/FMM approach for the PB equation adopts a nonadaptive tree structure for the sake of easy implementation, which is suitable for fairly uniform element distributions. For the surface integral equation formulation, as the elements distribute only on the surface of the molecule, at the lower levels of the tree structure a large number of boxes beyond the molecular surface are empty, which significantly compromises the computational efficiency of the algorithm because of the time and storage spent on these empty boxes. Moreover, the nonadaptive algorithm is more difficult to strike a load-balance between the number of elements in the local list (calculated directly) and those in the far-field (calculated using multipole and local expansions), thus further reducing its efficiency. By contrast, the adaptive FMM (AFMM) continues to subdivide boxes only until the number of elements in a box has reached a predefined number, thus creating a practically 'uniform' partition of particles in all childless boxes regardless of their sizes. In this paper, we present an improved implementation of the PB solver using an adaptive new version of FMM,[18] a "node-patch" discretization approach,[19] and the Krylov subspace iterative subroutines from the open source package SPARSKIT.[20] The resulting adaptive solver shows not only more efficient use of the memory but also significantly improves the load-balance between the local and far-field calculations, thus leading to faster calculation by several fold.

This paper is organized as follows. In Section 2, we describe the boundary integral equation formulation for the linearized PB solver. In Section 3, the "node-patch" discretization scheme is introduced to further reduce the number of unknowns. In Section 4, we discuss the Krylov subspace subroutines used in our solver, in particular, the package SPARSKIT and its convenient "reverse communication protocol". In Section 5, we briefly discuss the adaptive new version of FMM. In Section 6, numerical results are presented to benchmark the efficiency and accuracy of the solver, and finally in Section 7, we conclude this paper and discuss how to further optimize the solver using optimized oct-tree structure based on "spectral graph theory"[21] and parallelization on multicore multiprocessor computers.

2. Boundary Integral Equation Formulations

The Poisson−Boltzmann equation takes its most standard form as

$$-\nabla(\varepsilon\nabla\phi) + \kappa^2 \sin h(\phi) = \sum_{i=1}^{M} q_i \delta(r - r_i) \qquad (1)$$

When the electrostatic potential ϕ is small, the linearized PB equation can be obtained as

$$-\nabla(\varepsilon\nabla\phi) + \kappa^2\phi = \sum_{i=1}^{M} q_i \delta(r - r_i) \qquad (2)$$

When Green's second identity is applied, traditional boundary integral equations for the linearized PB equation for a single domain (molecule) can be written as,

$$\phi_p^{\text{int}} = \oint_s \left[G_{pt} \frac{\partial \phi_t^{\text{int}}}{\partial n} - \frac{\partial G_{pt}}{\partial n} \phi_t^{\text{int}} \right] \mathrm{d}S_t + \frac{1}{D_{\text{int}}} \sum_k q_k G_{pk} \qquad p, k \in \Omega \quad (3)$$

$$\phi_p^{\text{ext}} = \oint_s \left[-u_{pt} \frac{\partial \phi_t^{\text{ext}}}{\partial n} + \frac{\partial u_{pt}}{\partial n} \phi_t^{\text{ext}} \right] \mathrm{d}S_t \quad p \in \bar{\Omega} \qquad (4)$$

where ϕ_p^{int} is the interior potential at position p of the molecular domain Ω, q_k is the kth source charge, and $S = \partial\Omega$ is the molecular boundary. There are a variety of ways to specify the molecular boundary (solute−solvent dividing surface), and it is known that different specifications of the boundary can lead to very different results (see, e.g., ref 22). This is a practically important issue but is beyond the scope of this work. The particular surface types used in this work will be noted in the later sections when encountered. ϕ_p^{ext} is the exterior potential at position p, D_{int} is the interior dielectric constant, t is an arbitrary point on the boundary, and n is the outward normal vector at t. In the formulas, G_{pt} and u_{pt} are the fundamental solutions of the corresponding Poisson and Poisson−Boltzmann equations, respectively. When point p approaches the surface S, by satisfying the boundary conditions $\phi^{\text{int}} = \phi^{\text{ext}}$ and $D_{\text{int}}(\nabla\phi^{\text{int}} n) = D_{\text{ext}}(\nabla\phi^{\text{ext}} n)$, eqs 3 and 4 become a set of self-consistent boundary integral equations (denoted as nBIEs),

$$\alpha_p f_p = \oint_s PV\left[\varepsilon G_{pt} h_t - \frac{\partial G_{pt}}{\partial n} f_t \right] \mathrm{d}S_t + \frac{1}{D_{\text{int}}} \sum_k q_k G_{pk} \qquad p \in S \quad (5)$$

$$(1 - \alpha_p) f_p = \oint_s \mathrm{PV}\left[-u_{pt} h_t + \frac{\partial u_{pt}}{\partial n} f_t \right] \mathrm{d}S_t \; p \in S \qquad (6)$$

where PV denotes the principal value integral to avoid the singularity when $t \to p$, $f = \phi^{\text{ext}}$, $h = \nabla\phi^{\text{ext}} n$, and $\varepsilon = D_{\text{ext}}/D_{\text{int}}$. The coefficient constant α_p is 1/2 for a smooth surface, and more generally, it depends on the local surface

geometry at node p. For a vertex of a polyhedron, the coefficient α_p equals $A_p/4\pi$, where A_p is the interior solid angle at p. The constant of 1/2 has been usually used in previous BEM/PB work, while we have recently demonstrated that the use of a geometry-dependent coefficient significantly improves the overall numerical accuracy for the potential evaluation.[19]

The derivative BIEs (dBIEs) can be obtained by linearly combining eqs 5 and 6 and their derivative forms (for smooth surface case).

$$\left(\frac{1}{2\varepsilon}+\frac{1}{2}\right)f_p = \oint_s \mathrm{PV}\left[(G_{pt}-u_{pt})h_t - \left(\frac{1}{\varepsilon}\frac{\partial G_{pt}}{\partial n}-\frac{\partial u_{pt}}{\partial n}\right)f_t\right]\mathrm{d}S_t + \frac{1}{D_{\mathrm{ext}}}\sum_k q_k G_{pk} \quad p \in S \quad (7)$$

$$\left(\frac{1}{2\varepsilon}+\frac{1}{2}\right)h_p = \oint_s \mathrm{PV}\left[\left(\frac{\partial G_{pt}}{\partial n_0}-\frac{1}{\varepsilon}\frac{\partial u_{pt}}{\partial n_0}\right)h_t - \frac{1}{\varepsilon}\left(\frac{\partial^2 G_{pt}}{\partial n_0 \partial n}-\frac{\partial^2 u_{pt}}{\partial n_0 \partial n}\right)f_t\right]\mathrm{d}S_t + \frac{1}{D_{\mathrm{ext}}}\sum_k q_k \frac{\partial G_{pk}}{\partial n_0} \quad p \in S \quad (8)$$

where n is the unit normal vector at point t and n_0 is the unit normal vector at point p. The dBIEs lead to a well-conditioned (Fredholm second kind) system of algebraic equations. When Krylov subspace methods are applied to such systems, the number of iterations remains bounded even for a large number of elements. In our former work, we extended this form to systems of more than one separated molecules and provided a set of corresponding equations for force calculation.[6]

3. "Node-Patch" Discretization

After a typical triangular discretization, eqs 7 and 8 become

$$\left(\frac{1}{2\varepsilon}+\frac{1}{2}\right)f_p = \sum_t^T (A_{pt}h_t - B_{pt}f_t) + \frac{1}{D_{\mathrm{ext}}}\sum_k q_k G_{pk} \quad (9)$$

$$\left(\frac{1}{2\varepsilon}+\frac{1}{2}\right)h_p = \sum_t^T (C_{pt}h_t - D_{pt}f_t) + \frac{1}{D_{\mathrm{ext}}}\sum_k q_k \frac{\partial G_{pk}}{\partial n_0} \quad (10)$$

where T is the total number of discretized patches of the combined boundaries, while $2T$ represents the total unknowns of the system (i.e., f and h), and $\sum_k$ encompasses all the source charges of the system. The corresponding coefficient matrices are defined as follows:

$$A_{pt} = \int_{\Delta S_t}(G_{pt}-u_{pt})\mathrm{d}S, \quad B_{pt} = \int_{\Delta S_t}\left(\frac{1}{\varepsilon}\frac{\partial G_{pt}}{\partial n}-\frac{\partial u_{pt}}{\partial n}\right)\mathrm{d}S,$$

$$C_{pt} = \int_{\Delta S_t}\left(\frac{\partial G_{pt}}{\partial n_0}-\frac{1}{\varepsilon}\frac{\partial u_{pt}}{\partial n_0}\right)\mathrm{d}S,$$

$$D_{pt} = \int_{\Delta S_t}\frac{1}{\varepsilon}\left(\frac{\partial^2 G_{pt}}{\partial n_0 \partial n}-\frac{\partial^2 u_{pt}}{\partial n_0 \partial n}\right)\mathrm{d}S, \quad (11)$$

where the integrations are performed on the small patch ΔS_t. To obtain the above form, f and h are assumed to be constant in each ΔS_t patch. When p and t are nearby patches, eq 11 is performed by direct integration, otherwise the kernel for each patch integral is taken as a constant. The patch properties such as the normal vector and area are determined by the discretization method. A "node patch" discretization is employed in this work and will be described in the following paragraphs. In a typical iterative solution of the linear system eqs 9 and 10, the matrix-vector multiplication (first summations in eqs 9 and 10) needs to be performed in every iteration step, which accounts for the major computational cost. However, these computations can be conveniently accelerated by using FMM: for all the local pairs of p and t as defined in the FMM oct-tree structure, direct integration is performed over the corresponding patches, while the far-field calculation is achieved through the multipole expansion approximation. In addition, FMM is also used in the summation over all the source point charges as appeared in the last terms in eqs 9 and 10.

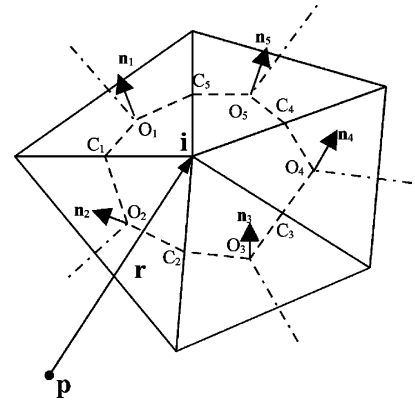


Figure 1. A "node patch" around the *i*th corner enclosed by the dashed lines is constructed on a triangular mesh. *O* and *n* are the centroid and normal vector of an element respectively, and *C* is the middle point of an edge.

There are two ways to treat the unknown f (or h) in the BEM approaches. The first is the so-called "constant element" approach which treats f (or h) as a constant on each element (face). Thus, the number of unknowns equals to the number of elements. Alternatively, f (or h) on each element can be obtained by linear interpolation from the unknowns of the three constituent nodes (also known as linear element approach), in which the number of unknowns equals the number of nodes. It is easy to show that, compared to the "constant element" approach, the "linear element" one leads to a reduction of the total number of unknowns by approximately a factor of 2, but a major disadvantage of the node-based approach lies in the introduction of additional complexity in its numerical implementation. In a recent communication, we introduced a node-patch scheme that appears to enjoy the benefits of both methods.[19]

The idea of the "node-patch" approach is to construct a "working" patch around each node instead of directly using the facet patch (element). We further assume that f (and h) is constant on this new "node-patch". A simple way to construct these new patches is illustrated in Figure 1, in which a "node patch" is constructed around the *i*th node that has five neighboring elements. The new patch is defined by the area encircled by a sets of points

$\{O_1, C_1, O_2, C_2, ...O_5, C_5, O_1\}$, where $\{O_l, l = 1, ..., 5\}$ are the centroids of the five adjacent triangles, and $\{C_l, l = 1, ..., 5\}$ are the midpoints of the five joint edges. It is easy to show that each triangular element contributes one-third of its area to the new "node-patch". Consequently, the far-field integrals on the new patch ΔS_i become

$$\int_{\Delta S_i} G_{pt} h_t dS \approx h_i G_{pi} \Delta S_i^a, \quad \Delta S_i^a = \frac{1}{3} \sum_{l \in \{L\}} \Delta S_l \quad (12)$$

$$\int_{\Delta S_i} \frac{\partial G_{pt}}{\partial n} f_t \mathrm{d}S \approx f_i G_{pi} \Delta S_i^b, \quad \Delta S_i^b = \frac{1}{3} \sum_{l \in \{L\}} \Delta S_l n_l \quad (13)$$

where n_l is the unit normal vector of the lth neighboring element, ΔS_l is the area of the lth adjacent triangular element, all the neighboring elements of the ith node form a set $\{L\}$, and ΔS_i^b should be considered as a vector. For near-patch integration, a normal quadrature method is used as in the constant element method. Similar treatments apply to the integrations for the kernel u and its derivative, as well as for the second-order derivative terms if the dBIEs are used.

There are three main advantages of this "node-patch" approach in BEM. First, as aforementioned, because of the reduction of the total number of unknowns when compared to the constant element method, the computational cost of solving the resulting linear system is accordingly reduced. The only additional computation is associated with the preprocessing of the geometric coefficients ΔS_i^a and ΔS_i^b in eqs 12 and 13. This, however, only constitutes a negligible portion of the total PBE solution time, and the geometric coefficients can also be saved for repeated use in iterative solving procedures. The second advantage, which is not so explicit, lies in the fact that relative to the linear element method, the "node-patch" method is significantly more efficient in searching and indexing the local list when used with any practical matrix storage format such as the Harwell−Boeing sparse matrix format or modified sparse column (row) format. Finally, in the "node-patch" method, the same as in the constant element method, the source and target are the same set of points, the nodes, which makes it straightforward to use any currently available FMM. Otherwise, if the source is different from the target as in the linear element method (where the convolution is done between two sets of data: the nodes and the quadrature points), still extra effort will be necessary to optimize the current FMM code to achieve comparable efficiency.

4. Krylov Subspace Methods

The discretized eqs 9 and 10 form a well-conditioned Fredholm second kind integral equation system, and a common practice for its efficient solution is to use Krylov subspace-based iterative methods. As the Fredholm second kind operator consists of an identity operator plus a compact operator whose eigenvalues only cluster at 0, it is well-known that the number of iterations in the Krylov subspace methods will be bounded, independent of the number of nodes in the discretization. Hence, the total number of operations required to solve eqs 9 and 10 is a constant (representing the number of iterations) times the amount of work required for a matrix vector multiplication. As will be discussed in next section, when the new version of fast multipole methods (FMM) are applied, the matrix vector product only requires $O(N)$ operations with an optimized prefactor; therefore, the linear equation system can be solved in asymptotically optimal $O(N)$ time.

Given an initial iterate x_0, the Krylov subspace method solves the linear system $Ax = b$ by iteratively minimizing some measure of error over the space $x_0 + \mathbf{K}_k$, where $\mathbf{K}_k = \mathrm{span}\{r_0, Ar_0, A^2r_0, ..., A^{k-1}r_0\}$ and r_0 is the initial residual usually defined as $r_0 = b - Ax_0$. On the basis of different measures of the error and different types of matrices, there are many different implementations of the Krylov subspace method. As the matrix A in eqs 9 and 10 is nonsymmetric and there is no fast algorithm for multiplying the transpose of A with an arbitrary vector, in our solver, we have tested four different Krylov iterative subroutines from the open source package SPARSKIT developed by Saad and collaborators.[20,23] These are the full GMRES, the restarted GMRES, the biconjugate gradients stabilized (BiCGStab) method, and the transpose free QMR (TFQMR). Our preliminary numerical experiments show that all these solvers perform well in most cases, and not surprisingly, the full GMRES seems to perform the best. Also, as will be shown in Section 6, the number of iterations using the iterative solvers in SPARSKIT is often less than the numbers we observed in our previous implementations in which a different Krylov subspace package is used, partly because of our optimized initial guess and a few other improvements in the present code.

An interesting feature of the iterative solvers in SPARSKIT is the so-called "reverse communication protocol" (confer the ITSOL directory in SPARSKIT[20]), which avoids having to call the matrix vector product subroutine from inside the Krylov solver. Instead, the Krylov solver provides a vector and asks for the matrix vector multiplication result as future input. Therefore, it is unnecessary to pass the parameters in the FMM subroutines to the Krylov solver, which means easier interface between the FMM and SPARSKIT, and easier memory management.

5. Adaptive Fast Multipole Methods

The fast multipole method was first invented by Greengard and Rokhlin in 1987[13] for the fast evaluation of the Coulomb interactions of N particles. For any given cluster of particles, as the far-field potential due to these particles is a smooth function and can be represented by a few terms of spherical harmonic expansions in 3D (the Laurent expansions in 2D), the interactions can therefore be efficiently accounted for using a divide-and-conquer strategy as follows: first, an oct-tree structure is generated so each childless box (leaf node) only contains a few particles; next, an upward sweep is executed to form the multipole expansions which carry the far-field information for all boxes, by using the particle information directly for the childless boxes, and by shifting the children's multipole expansions to parent level boxes (multipole-to-multipole); third, a downward sweep is used for each box to gather far-field information which is stored in a local expansion. At each level, the box first obtains very far-field information from its parent using a local-to-local

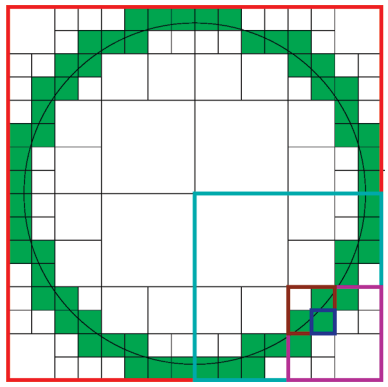


Figure 2. A schematic 2D adaptive tree structure.

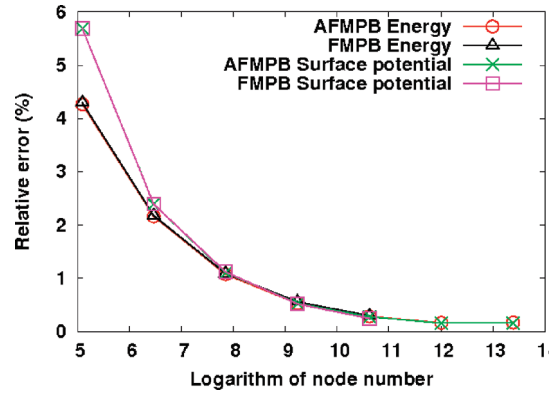


Figure 3. Accuracy of energy and potential calculations with the conventional and adaptive solvers. The relative errors of surface potentials are averaged over all node points.

translation and then shifts the multipole expansions from the "interaction list" (far-field boxes of itself which are not far field of its parent) to its local expansion (multipole-to-local); fourth, the local expansions of the childless boxes are evaluated at each particle location to account for all the far-field particle interactions; finally, the local particle interactions are evaluated directly. Notice that the number of boxes is $O(N)$ and the amount of work for local direction interactions is also $O(N)$; therefore, the algorithm is asymptotically optimal $O(N)$.

However, because of the large number of boxes in the interaction list and $O(p^2)$ operations for each multipole-to-local translation when p terms are used in the expansion, many numerical implementations reveal that the 1987 version of FMM is less competitive compared with other methods including the PME[9] and tree code,[12,24] and the prefactor in $O(N)$ is often >10 000. To further accelerate its performance, in 1997, a new version of FMM was presented by Greengard and Rokhlin for the Laplace equation,[14] in which exponential expansions are introduced to diagonalize the multipole-to-local translations, and a "merge-and-shift" technique is used to further reduce the number of such translations. Numerical experiments show that the new version of FMM breaks even with direct calculation when the number of particles $n = 500$ for three digit accuracy, and $n = 1000$ for six digits for Coulomb interactions. In our previous work,[6] the new version of FMM was implemented for the Laplace and linearized PB equations for the efficient calculation of electrostatic interactions. As far as we know, this was the only LPB solver using the new version of FMM.

In this paper, we further improve our solver by using an adaptive new version of FMM. Unlike our previous implementation where a uniform oct-tree is generated, we remove those empty nodes in the oct-tree structure by only subdividing a box when the number of particles it contains is more than a prescribed number. Notice that this is important as all the unknowns are only located on the surface of the molecule. Figure 2 schematically shows a 2D adaptive tree structure for a circular boundary problem. There would be a total number of 256 smallest boxes when using 4 levels of box subdivisions in the uniform quad-tree structure, while only 64 boxes on the circular boundary are counted for in the adaptive tree structure. As shown by our preliminary numerical results in next section, the adaptive tree structure not only improves the efficiency of the code but also reduces the required memory storage so larger problems can be computed.

Instead of discussing technical details of the adaptive new version of FMM, we refer the readers to existing literatures. The new version of FMM was introduced in ref 14 for the Laplace equation, the corresponding adaptive version was discussed in ref 16, the new version of FMM for the linearized PB equation (also called Yukawa or modified Helmholtz equation) was discussed in 17, and our LPB solver using a uniform oct-tree structure was presented in ref 6.

6. Benchmarks

A. A Spherical Cavity. The first system selected is a point charge located at the center of a spherical cavity. We examine the accuracy of the algorithm at different discretization resolution by comparing the calculated energy and potential to those from the analytical solution. We first note that for any relatively uniform particle distribution, AFMM and FMM maintain almost the same level of accuracy because the error is largely dependent on the box level and the number of terms used for expansion but not on how the data structure is represented, i.e., an adaptive vs nonadaptive tree structure. This is confirmed by the results displayed in Figure 3 that for both the energy and potential calculations the data lines obtained with AFMM overlap with those with FMM. It is also worth noting that the energy and potential calculations show similar errors, both of which reduce roughly linearly as the element size decreases. The energy and potential calculations converge with a relative error <0.2% when the surface mesh resolution is finer than 0.25 Å^2 (the surface area of a single triangular panel). At any given mesh resolution, the numerical error is bounded with the FMM (for far-field calculation) and the local numerical integration, but how the resolution and quality of meshes might affect the accuracy of the calculation is somewhat more difficult to quantify. While very good converging behavior (and well-defined converging resolution) is observed here for a spherical case, the calculations are performed on a single charge with a perfect geometry. Therefore, further studies would be necessary to assess the convergence criterion for more complex biological systems.

Table 1. Performance Comparison on a Spherical Cavity Case at Different Level of Discretization Resolution

number of elements	CPU (s)		memory (megabytes)		max. level	
	AFMM	FMM	AFMM	FMM	AFMM	FMM
162	0.05	0.13	2.7	2.7	3	2
642	0.21	0.62	7.0	7.9	4	3
2562	0.89	2.66	24.4	54.0	5	3
10242	4.63	11.44	113.3	241.0	6	4
40962	19.26	57.73	511.8	935.0	7	5
163842	78.35	-	2152.1	-	8	-
655362	1051.20	-	7900.7	-	9	-

A particular advantage of the adaptive algorithm, when compared to the nonadaptive one, is its lower memory usage. This is due to the fact that in a nonadaptive tree structure, when the elements only distribute on the surface as in BEM, a large number of boxes beyond the molecular surface are empty, leading to unnecessary memory usage for storing these empty boxes and their associated expansion coefficients. By contrast, the adaptive FMM continues to subdivide boxes only until the number of elements in a box has reached a predefined number, thus creating a practically 'uniform' partition of particles in all childless boxes regardless of their sizes. In our PB solver, the memory taken up by the FMM part constitutes a considerable part of total memory usage. But to what extent depends on how much information for local-field calculations the BEM saves during solution of the PBE. We here estimated the memory usage in a stand-alone FMM code for a test case where all the particles uniformly distribute on a spherical surface. Our results show a memory reduction of >10 fold with a 5-level-subdivision of the box, or more generally a reduction of $\sim 2^{n-1}$ fold when level n is greater than 6. In any real PB calculation, the above simplified analysis is not valid anymore because of several contributing factors, such as nonideal shape and/or charge distribution of the system, and additional memory usage by the other part of the program, but the overall trend remains evident that the adaptive algorithm still uses less memory than that of the nonadaptive one (Table 1). The comparison becomes even more favorable toward the adaptive solver when the subdivision becomes finer and/or the system size becomes larger. Because of this improvement, the PB calculations can now be performed on our desktop computer for systems with much more surface elements (e.g., 163 842 and 655 362 in Table 1) than what can be handled previously by the nonadaptive solver. For both solvers, the node-patch approach is used.

Furthermore, the adaptive FMM can strike a better load-balance for treating elements in the local list (calculated directly) and those in the far-field (calculated using expansion coefficients), while in the nonadaptive algorithm the partition between the local and far-field elements is greatly limited by the power growth of memory ($\sim 8^n$) as the number of levels n increases in an oct-tree data structure. As shown in Table 1, with 655 362 surface elements, the adaptive solver can handle a maximum tree level of 9 without causing any memory overflow problem on our 8-gigabyte desktop computer. However, for the nonadaptive algorithm, the maximum level that can be handled is only 6 (data not shown). Further tests reveal that level 9 enables the most balanced local and far-field calculations, thus is optimal for this particular case. When the calculation is otherwise performed at level 6, the FMM calculation is very unbalanced in a sense that too many elements are assigned to the local list for direct calculation. Specifically, the direct calculation takes more than 90% of the total computing time, while the far-field part takes less than 5%. The poor load-balance significantly compromises the overall efficiency of the calculation. For most of the calculations, an average speedup of 2−3 fold has been observed by using the adaptive algorithm, while better performance is generally expected for larger systems. Therefore, as compared to the nonadaptive solver, our new adaptive implementation not only makes more efficient use of the memory but also increases the calculation speed quite significantly.

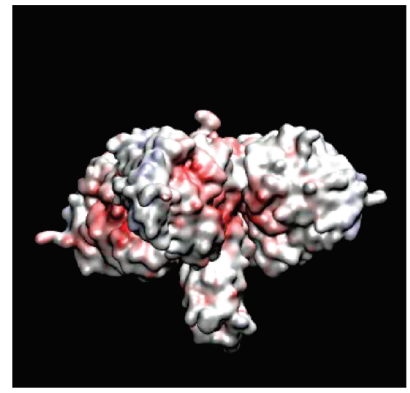


Figure 4. Electrostatic potential surface of the acetylcholinestase.

Table 2. Performance Comparison on the Acetylcholinesterase Tetramer

	old FMPB	AFMPB
solvation energy (kcal/mol)	−8341.3	−8342.4
CPU time (s)	695.5	94.2
memory (gigabytes)	1.40	1.05
max. level	6	9
number of iteration	18	15

B. Acetylcholinesterase Tetramer. As a representative protein system, we chose the acetylcholinesterase tetramer, which contains 36 638 atoms with a dimension of 135 Å × 112 Å × 115 Å (Figure 4). The molecular surface (also known as the solvent-excluded surface), which is the surface traced by the inward-facing surface of the probe sphere, is used as the boundary. The surface discretization using MSMS[26] resulted in 124 254 triangular elements and 62 095 nodes. Both the adaptive and nonadaptive solvers can handle this system well on a typical 2008 desktop computer, giving very comparable solvation energy (see Table 2). It would have seemed surprising at the first sight that the adaptive solver uses almost the same amount of memory as the nonadaptive one, but when you note that far more levels of box subdivision are used by the adaptive solver than the nonadaptive one, the seemingly conflicting results are actually easy to understand. Because of its ability to involve more tree levels, the new adaptive solver runs about 7 times faster than the old one, one of the greatest speedups observed in all our test calculations. The main reason for the observed

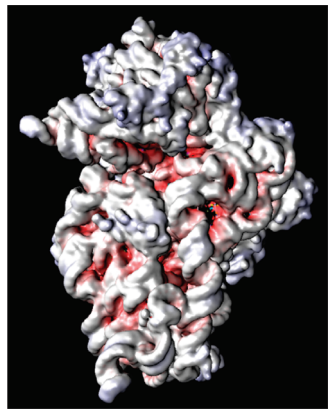


Figure 5. Electrostatic potential surface of the 30S ribosome subunit.

acceleration as aforementioned is because the new solver can have a more balanced work load for both the local and far-field computation. By contrast, the old nonadaptive solver cannot run with more than 6 levels, thus leading to a too heavy local computation load. Another acceleration factor comes from the reduction of the iterative steps (Table 2) in solving the linear system by using SPARSKIT as analyzed above.

C. 30S Ribosome Subunit. Finally, we show the results of a 30S ribosome system (PDB code: 1fjf[25]), which is nontrivial to compute with other or our earlier PB solvers using a serial version on a desktop, because of its far greater memory requirement. The 30S ribosome subunit consists of 21 peptides and a 1540-nucleotide RNA subunit, with a total of 88 431 atoms (including hydrogen atoms) and a dimension of 211 Å × 177 Å × 200 Å (Figure 5). Because of its size, the software MSMS fails to generate a molecular surface mesh for ribosome. So, the Gaussian surface, which is a level-set of the summation of the spherically symmetrical Gaussian density distribution centered at each atom of the molecular, is used as the molecular boundary, and the surface discretization is performed using the software Gamer.[27] The surface discretization results in 343 028 triangular elements and 171 288 nodes. The edge length resolution is about 1 Å. The whole computation takes ∼21 min on our desktop machine (Intel(R) Xeon(TM) CPU 3.00 GHz, 4GB memory), with a memory usage of 2.6 GB. An 11-level tree structure is used for optimal efficiency, and 92 iteration steps are taken to obtain a converged solution. Figure 5 shows the computed electrostatic potentials mapped on the molecular surfaces of the 30S ribosome.

7. Conclusions

We have described a new implementation of our BIE-based PB solver that uses an adaptive FMM for accelerating the *N*-body type surface integration. Other salient points of our current implementation include (1) the well-conditioned formulation that is extended to multidomain systems, (2) the development of an efficient patch-node scheme for surface discretization, and (3) efficient use of Krylov subspace method for iterative solution of the linear system. The adaptive FMM reduces the total number of boxes by using nonuniform oct-tree structure and thus leads to significant reduction in memory usage. Because of its ability to keep a better load-balance for the local and far-field calculations, the speedup is also quite significant when compared to our earlier version of the nonadaptive solver.

The resulting solver was tested with several applications. The accuracy of the new algorithm was first examined by direct comparison with the analytical solution of a point charge located at the center of a spherical cavity. It is found that the solvation energy of our spherical cavity with radius 50 Å converges with a relative error <0.2% when the surface mesh resolution is below 0.25 Å^2 of each triangular element. Our new PB solver significantly outperforms our earlier nonadaptive solver and shows a stronger linear growth of both memory and computational cost with the number of unknowns. Primarily because of more efficient memory allocation, the new solver enables very large-scale calculation to be executed on a typical 2008 desktop machine. A PB calculation on the 30S ribosome (88 431 atoms) illustrated the capability of the code. The new solver is also very suitable for doing calculation on large-scale biomolecular assembly or complex systems that comprise a set of molecules with large separations between them, though we do not show any test calculation for such cases. Further applications of the methodology are in progress, including the coupling with molecular dynamics/Brownian dynamics simulation and other continuum models for studying molecular interactions and dynamics of biological systems.

However, in order to perform dynamics simulation or study other electrostatics-controlled dynamical process, our code needs further improvement. To do this, our current efforts include (1) generating an optimized oct-tree structure using spectral graph theory,[21] and (2) parallel implementation of the code on computers with shared and/or distributed memory. Another important direction is to come up with a more efficient way to generate molecular surfaces, which seems to be the current bottleneck for performing a fully dynamical simulation (PB solution at every time step). Finally, the surface specification itself is an important and open issue as aforementioned. As the general practice in BEM, this work uses a single surface separating the solute and the solvent. On the other hand, in many finite-difference methods, a second surface, the so-called Stern layer, is introduced to account for the fact there is a layer in the solvent to which mobile ions cannot access. Complicated by some other factors (parameters) in the setup of a PB calculation that can also affect the final results, it is hard to conclude thus far which surface specification is the best. Likewise, the surface model adopted in BEM will need further tests and comparisons with experiments or other more accurate computations.

Acknowledgment. We thank many of our colleagues and collaborators for their contributions and suggestions. In particular, our code uses many existing open source codes, including the SPARSKIT by Saad and collaborators,[20] MSMS,[26] and Gamer[27] for mesh generation, VMD[28] for visualization, and several important subroutines in the new version of FMM from Profs. Greengard and Rokhlin's group. This work was supported by HHMI, NIH, NSF (J.H.: NSF0811130 and NSF0411920, J.A.M.: MCB0506593), and

the NSF Center of Theoretical Biological Physics (CTBP). B.Z. is partially funded by the "100 Talents Projects" of the Chinese Academy of Sciences, China. X.C. is partially funded by the Computer Science and Mathematics Division at Oak Ridge National Laboratory. Their support is thankfully acknowledged.

References

(1) Baker, N. A. *Methods Enzymol.* **2004**, *383*, 94.

(2) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.

(3) Holst, M.; Baker, N. A.; Wang, F. *J. Comput. Chem.* **2000**, *21*, 1319.

(4) Boschitsch, A. H.; Fenley, M. O.; Zhou, H. X. *J. Phys. Chem. B* **2002**, *106*, 2741.

(5) Kuo, S.; Altman, M.; Bardhan, J.; Tidor, B.; White, J. Fast methods for simulation of biomolecules electrostatics. Proceedings of the IEEE/ACM International Conference on Computer Aided Design; San Jose, CA, November 10–14, 2002, p 466.

(6) Lu, B.; Cheng, X.; Huang, J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (51), 19314.

(7) Phillips, J. R.; White, J. A Precorrected-FFT Method for Capacitance Extraction of Complicated 3-D Structures. International Conference on Computer-Aided Design; San Jose, CA, November 6–10, 1994.

(8) White, J.; Phillips, J. R.; Korsmeyer, T. Comparing Precorrected-FFT and Fast Multipole Algorithms for Solving Three-dimensional Potential Integral Equations. Proceedings of the Colorado Conference on Iterative Methods; Breckenridge, CO, April 4–10, 1994.

(9) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(10) Lee, H.; Darden, T.; Pedersen, L. *Chem. Phys. Lett.* **1995**, *243* (3−4), 229.

(11) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions*; Dover, New York, 1970.

(12) Appel, A. W. *SIAM J. Sci. Stat. Comput.* **1985**, (6), 85.

(13) Greengard, L.; Rokhlin, V. *J. Comp. Phys.* **1987**, *73* (2), 325.

(14) Greengard, L.; Rokhlin, V. *Acta Numer.* **1997**, *6*, 229.

(15) Greengard, L.; Wandzura, S. *IEEE Comput. Sci. Eng.* **1998**, *5* (3), 16.

(16) Cheng, H.; Greengard, L.; Rokhlin, V. *J. Comp. Phys.* **1999**, *155* (2), 468.

(17) Greengard, L.; Huang, J. F. *J. Comp. Phys.* **2002**, *180* (2), 642.

(18) Fast Multipole Methods (Beta). http://www.fastmultipole.org/ (accessed Dec 14, 2008).

(19) Lu, B.; McCammon, J. A. *J. Chem. Theory Comput.* **2007**, *3*, 1134.

(20) SPARSKIT. http://www-users.cs.umn.edu/∼saad/software/SPARSKIT/sparskit.html (accessed Mar 8, 2005).

(21) Chung, F. R. K. Spectral Graph Theory; Published for the Conference Board of the mathematical sciences by the American Mathematical Society, Providence, RI, 1997.

(22) Dong, F.; Vijayakumar, M.; Zhou, H. X. *Biophys. J.* **2003**, *85* (1), 49.

(23) Saad, Y.; Schultz, M. *SIAM J. Sci. Statist. Comput.* **1986**, *7*, 856.

(24) Barnes, J.; Hut, P. *Nature* **1986**, *324*, 446.

(25) Wimberly, B. T.; Brodersen, D. E.; Clemons Jr., W. M.; Morgan-Warren, R.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327.

(26) MSMS. http://www.scripps.edu/∼sanner/html/msms_home.html (accessed Feb 5, 1996).

(27) Yu, Z.; Holst, M.; Cheng, Y.; McCammon, J. A. *J. Mol. Graph. Model.* **2008**, *26*, 1370.

(28) VMD. http://www.ks.uiuc.edu/Research/vmd/ (accessed Mar 20, 2006).

CT900083K

JCTC Journal of Chemical Theory and Computation

Analyzing the Selectivity and Successiveness of a Two-Electron Capture on a Multiply Disulfide-Linked Protein

Élise Dumont,*,† Adèle D. Laurent,‡ Pierre-François Loos,§,‡ and Xavier Assfeld‡

Laboratoire de Chimie, UMR 5182 CNRS École Normale Supérieure de Lyon, 46, allée d'Italie, 69364 Lyon Cedex 07, France, and Équipe de Chimie et Biochimie Théoriques, UMR 7565 CNRS-UHP, Institut Jean Barriol (FR CNRS 2843), Faculté des Sciences et Techniques, Nancy-Université, B.P. 70239, 54506 Vandoeuvre-lès-Nancy, France

Received February 23, 2009

Abstract: Hybrid QM/MM calculations were performed on a circular macropeptide (kalata B1, PDB ID 1NB1) containing three disulfide linkages, to evaluate their respective reactivities toward (gas-phase) electron valence-attachment of one and two electron(s). The three disulfide bonds -CH_2-S-S-CH_2- were simultaneously described at the MP2/6-31+G**(S),6-31G*(C,H) level of theory, and the remaining of the 29 residues of kalata B1 were described by the CHARMM27 force field. The one-electron addition is favored on the linkage between cysteine residues 1 and 15, Cys(1−15), by ca. 1 eV over the two other disulfide linkages. The decomposition of the overall effect into geometrical and electrostatic contributions evidence (i) the key role of an arginine (R24) and (ii) a weaker geometrical penalty for elongating the nonstructural Cys(1−15) linkage. The addition of a second electron leads to the formation of the dithiolate Cys(1,15), favored by more than 1 eV over other adducts (either dithiolates or diradical dianionic species). This can be traced back to a structural reorganization, with a flip of R24 side chain. Its positively charged extremity points almost equidistantly toward one thiolate -CH_2-S^-, hence stabilizing this dianion.

I. Introduction

The existence of three-electron two-center (2c-3e) bonds has been postulated by Pauling[1] as early as 1931. An elegant theory was derived five decades later for predicting the relative stability of such hemibonded species[2] and was closely related to experimental data.[3] Their stability has been proved by a wide range of technics (pulse radiolysis,[4−6] electron spin resonance,[7] laser flash photolysis,[4] electrochemistry),[8] with a typical dissociation energy (ca. 20−30 kcal/mol) allowing a proper observation.

A strong motivation for the study of 2c-3e systems lies in their importance in reactivity of biological systems. For instance, they serve as 'relay stations'[9] in ubiquitous electron transfers.[10,11] Special importance is given to disulfides, because of their essential role for structure and reactivity of proteins. These radical anions (noted 2S-3e) have thus been intensively studied, either on model organic compounds,[12−16] organometallic complexes,[17,18] and biological systems, in which they have been recognized as long-lifetime intermediates.[19,20]

One would like to gain insight into the factors governing the formation, the stability and the outcome of these transient 2S-3e intermediates in a complex environment. Recently enough, Weik and co-workers have nicely demonstrated using X-ray synchrotron radiations the high specificity of low-energy electrons addition,[21] with a valence attachment on low-lying σ^* (SS) orbitals. Quantum calculations, alongside with topological analysis,[22] offer a complementary view, often more quantitative, on the structure and reactivity of the 2S-3e intermediates. Many questions remain answerless

* Corresponding author e-mail: elise.dumont@ens-lyon.fr.
† UMR 5182 CNRS École Normale Supérieure de Lyon.
‡ Nancy-Université.
§ Present address: Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia.

10.1021/ct900093h CCC: $40.75
Published on Web 04/30/2009

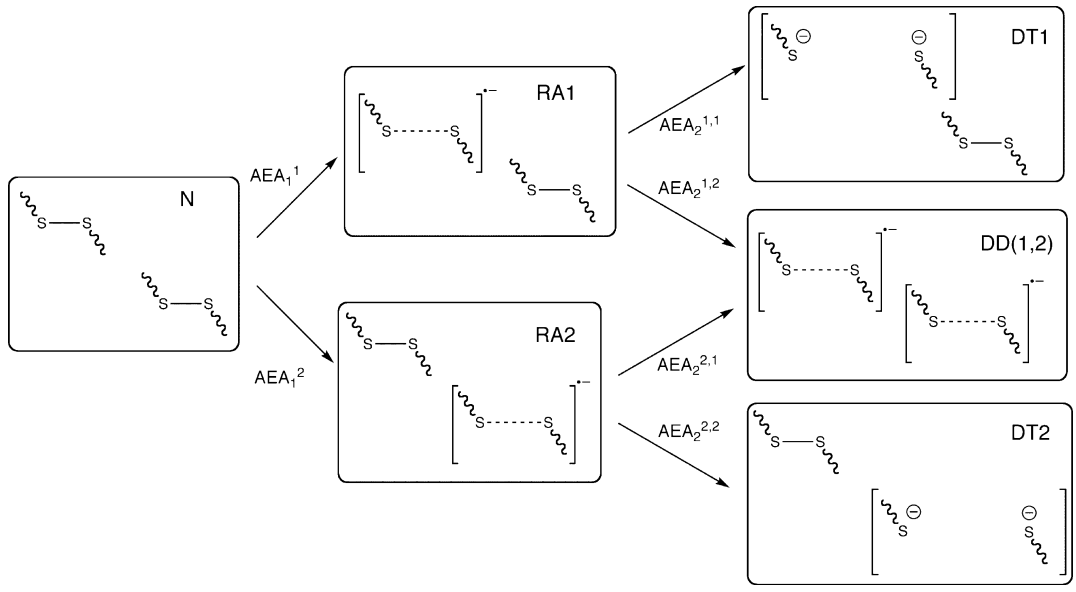


Figure 1. Schematic view of the possible outcomes of the stepwise two-electrons addition on a two-disulfide linked system. The square box represents the proteinic environment. The inner competition between the two disulfide bonds for the first electron uptake, with formation of radical anions (RA), is addressed by computing the respective adiabatic electron affinities AEA_1. Similarly, the addition of a second electron can form either a dithiolate (two possibilities DT1 or DT2) or a diradical dianion noted DD(1,2). Relative energies are computed to gain insights on such competitions.

concerning their formation, which is directly quantified by the electron affinity (EA). This intriguing reaction is formally simple and presents two key characteristics (cf. Figure 1): the drastic disulfide lengthening (by ca. 0.7 Å) and the charge difference between the reactant and the product. This gives rise to two major contributions, geometric and electrostatic, that both impact EA. Striking examples have been reported for the huge modulation by the following:

1. the conformational strain or topological frustration that strongly favors an electron uptake.[23−25] This purely geometric effect enhances EA by ca. 1 eV for a Cys-Gly-Pro-Cys motif (CGPC), which forms the active site of Trxh1, an antioxydant enzyme from the thioredoxin superfamily.
2. the secondary structure, for instance the effect of a α-helix dipole (+0.9 eV for an Ala_{12} grafted on CGPC, constituting a peptidomimetic for Trxh1),[26]
3. a point charge of +1 au even at a distance of 10 Å[27] or, more realistically, a charged residue in the vicinity of a disulfide (accounting for ca. 2.0 eV from the Lys40 residue of Trxh1).[28]

These simple considerations usually suffice to conclude on the relative reactivity of two highly similar disulfide bridges, e.g. to discuss mutation effects. Indeed, most of the available results so far have focused on Trx enzymes possessing a unique, highly reactive, disulfide linkage.[24,28] But several other questions naturally arise as a mutiply disulfide-linked protein is considered, e.g. *Torpedo Acetylcholine Esterase* (TAchE) in the original paper by Weik and co-workers.[21]

The first question concerns the relative order of reactivity of disulfides, with an inner competition to treat. Redox reactions do not involve a flow of electrons but rather one (or two), whose attachment is highly specific.

Other questions arise for the addition of a second electron. At first sight, the beautiful X-ray structure of irradiated *TAchE*, with each of its three disulfides in radical anionic form, may suggest that *n successive* electron additions on *n* disulfide linkages results in the formation of *n* radical anions. But the electron uptake could also occur on a 2S-3e bond, forming a dithiolate, especially in solution with no packing effects. Such a cleavage results in a fragmentation of the protein. Calculations provide a reliable way to gain some insights on the electronic pathway (inner competition, cf. Figure 1) for the second EA, while no information can be inferred from experimental data as *all* disulfide bridges are inevitably damaged under radiation process. For instance, quantum mechanics (QM) calculations on the isolated active site of *TAchE* prove the constrasted disulfide reactivity.[20]

In this study, we have undertaken a systematic study of two successive electron attachments on a small circular macropeptide, widely studied in the literature, kalata B1 (kB1). It is the prototype of the cyclotide family, small disulfide rich macropeptides isolated from plants. The three-dimensional structure (Figure 2) is well-defined with a cyclic backbone (Mobius type with a *cis* proline) and three interlocking disulfide linkages, forming a highly characteristic cystine knot motif. The latter not only maintains the circular compact folding (thermal stability) but also enhances disulfide reactivity because of the constrained topology. A whole line of research now consists in tuning in a controlled way infectiologic properties of cyclotides (HIV inhibitors,[29] antimicrobial).[30]

Some of the proper characteristics of cyclotides make them perfect candidates, in the context of this study, with several advantages over other systems, notably the following: three disulfides linkages at first sight rather similar, a circular structure that bypasses the need to cap the N- and C-terminal

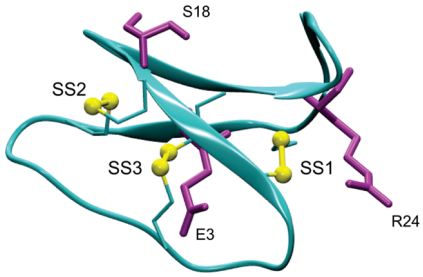


Figure 2. Cartoon representation of kalata B1 (PDB ID 1NB1). No proper secondary structure is defined because of the sequence short size and the cystine knot motif (interlocking arrangement of the three disulfide linkages Cys(1−15), Cys(10−22), and Cys(5−17) — labeled SS1, SS2, and SS3 on this scheme and represented with (yellow) balls). The latter imposes a tightly bent cyclic structure (backbone in cyan), with a Mobius topology (*cis* proline). Arginine R24, serine S18, and glutamate E3 side chains strongly tune disulfide electron affinities and are labeled and depicted with (purple) sticks.

residues, and, first and foremost, a wealth of experimental information.[31−34]

We built up in recent works[23,24,26] a methodology specifically tailored to accurately describe electron attachment on disulfide-linked systems, which is recalled in Section II. The selectivity of the first one-electron addition (inner competition) is analyzed with three different steps in Subsections 3.1, 3.2, and 3.3. The addition of a second electron is treated in the last Subsection (3.4); all possible adducts (nine) are considered to identify the electronically most stable product.

II. Computational Methodology: QM/MM Scheme

Due to the relatively large size of kB1 (29 residues, 376 atoms) and the high level of theory needed for describing electron attachment on disulfide bonds, hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods offer a near-optimal approach. Moreover, they enable a decomposition of the overall EA into geometric and electrostatic contributions, as detailed in the last Subsection.

A. QM Description of 2S-3e Bond and Definitions of Relative AEAs. Explicit treatment of electron correlation is essential for an accurate description odd-electrons bonds. Second-order Møller−Plesset perturbation theory (MP2)[35] has proved its reliability for a proper description of 2S-3e bonds.[36,37] First adiabatic electronic affinities (AEA_1) of optimized structures were defined, as usual, as the difference between energies of the optimized reactant (neutral compound, N) and product (radical anion, RA):

$$AEA_1 = E(N) - E(RA) \quad (1)$$

AEAs have proved to be highly sensitive to the basis set,[38] which needs to be carefully calibrated to treat neutral and anionic species on the same footing. In contrast, relative values ΔAEA are stable as soon as the basis set includes diffuse functions on the sulfur atoms — one benefits from a cancelation of errors.[24,25] They were defined with respect to a L,L-cystine capped by acetyl and N-methylamide (cf. Figure 3), which we chose as a reference (cf. eq 2).

$$\Delta AEA = AEA_{kB1} - AEA_{L,L-cys} \quad (2)$$

In this study, we chose a mixed Pople basis set, with 6-31+G** on sulfur and 6-31G* for carbon and hydrogen atoms.

B. Two-Layers Partioning of a Disulfide-Linked Peptide: QM/MM Scheme and Classical MM Description. A double proximal C_α–C_β frontier is defined, isolating the -CH_2-S-S-CH_2- fragments of the three cystines (cf. Figure 3), within a hydrogen link-atom (HLA) scheme.[39,40] The scaling factor corresponding to the ratio between $R(C_\beta - HLA)$ and $R(C_\alpha - C_\beta)$ is fixed to 0.71.

The MM surrounding is described with the CHARMM force field using the CHARMM27 parameters for proteins.[41−43] The van der Waals parameters of the QM atoms are set to the values defined for the corresponding atom type of the force field. To avoid an overpolarization of the C_β-HLA bonds, the nearby C_α point charge, q_{C_α}, initially equal to 0.07 e, has been set to zero. The overall electroneutrality of the MM part is ensured by a redistribution on the nitrogen (−0.47 → −0.435 e) and carbon (0.51 → 0.545 e) neighboring atoms — cf. Figure 3. We checked on L,L-cystine and diethyldisulfide[26] that this operation does not impact relative electron affinities. The placement of a frontier along a covalent bond inevitably introduces an artifact. But, we have recently discussed the frontier effects on a model compound (diethyldisulfide)[24] and established the stability of relative energies with respect to the level of theory.

In this study, the high-level QM part corresponds to the union of the three -CH_2-S-S-CH_2- fragments, which defines a global wave function. Very similar geometric and energetic data for the additions of (i) one electron and of (ii) a second one on the same linkage (formation of a dithiolate) are obtained whether the QM part is limited to a single cystine

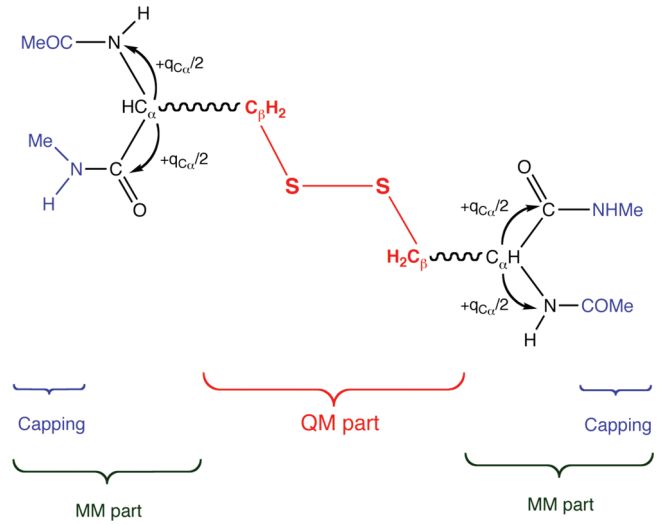


Figure 3. QM/MM partition adopted for describing electron addition on a disulfide linkage, illustrated on the capped L,L-cystine. This prototypical peptide constitutes the reference compound in this study. Wavy lines denote the C_α–C_β bonds which have been defined as QM/MM frontiers in this work. Atoms in bold (red) are treated with the MP2 method. Arrows indicate the charge redistribution ensuring the electroneutrality of the system. Terminal capping groups (NHMe and COMe) are indicated in blue.

Table 1. Geometrical Parameters (Respectively Bond Lengths, Bending and Dihedral Angles) and First Adiabatic Electron Affinities AEA_1 of kB1[a]

			structure				electron affinities			
linkage	label		*d*(S-S)	∠(S-S-C)	τ(C-S-S-C)	rmsd	AEA_1	ΔAEA_1	ΔEA_1^{elec}	ΔAEA_1^{geom}
Cys(1−15)	SS1	N	2.05	103.0, 106.0	95.5		1.58	0.52	0.78	−0.26
		RA	2.75	94.1, 106.8	132.5	0.87				
Cys(5−17)	SS2	N	2.06	102.8, 101.7	76.4		0.68	−0.38	0.31	−0.69
		RA	2.77	97.5, 91.7	70.3	0.23				
Cys(10−22)	SS3	N	2.06	101.3, 102.9	106.6		0.51	−0.55	0.24	−0.79
		RA	2.83	103.6, 104.4	81.7	0.28				
L,L-cystine	—	N	2.05	102.7, 104.7	75.2		1.06			
	—	RA	2.80	95.3, 95.8	66.5					

[a] The level of theory is detailled in the text. Absolute and relative values ΔAEA, with respect to L,L-cystine, are given in eV, for each of the three disulfide linkages. They are decomposed into electrostatic ΔAEA^{elec} and geometric ΔAEA^{geom} contributions. N or RA refer to neutral or radical anionic species. RMSD are reported for each RA with respect to the common neutral reference N.

or encompasses the three cystinyl fragments — cf. Table 2, Supporting Information, showing that having multiple QM/MM boundaries (six) does not induce any additional error.

Hybrid QM/MM calculations were performed with a modified version of the Gaussian 03 series of programs[44] linked to the Tinker software[45] for the MM calculations. Final geometrical parameters are given in angstroms (Å) and degrees. rmsd between neutral and (di)anionic forms were computed following the method of Kabsch[46] as implemented in the VMD software[47] — hydrogen atoms were excluded. The keyword guess = alter was used to force the initial SCF guess, thus obtaining each specific localized radical anions (RA), diradical dianions (DD), or dithiolates (DT). No spin contamination was observed for RA, with values of $\langle S^2 \rangle$ never greater than 0.77 (to be compared to the exact value of 0.75). DD can be found either in the triplet or the singlet states. For triplet states, $\langle S^2 \rangle$ values were never greater than 2.03 (to be compared to the exact values of 2.00), such that, again, no contamination spin will affect our results. The latter is observed for singlet states ($\langle S^2 \rangle$ up to 1.04) but does not affect the energetic results since singlet–triplet energy difference is negligeable (less than 1 kcal/mol, systematically in favor of the triplet state) was observed.

Cartesian coordinates for the NMR solution structure of kalata B1 were employed (PDB ID 1NB1). Each of the 20 experimental lowest energy geometries lead to the same 3D structure after classical optimization. Classical preoptimizations were performed using the minimize procedure of the TINKER suite of programs, with the lowest convergence criterion implemented — rms gradient of 0.1 kcal/mol/Å. All 20 NMR structures provide the same geometry of neutral kalata B1 (rmsd ranging between 0.257 and 0.332 with respect to PDB initial geometries). Starting from this structure of neutral kB1, QM/MM optimizations were performed for each electronic state (N, RA, DT, or DD): all residues (backbone and side chains) were varied, and the convergence was tested against standard criteria of Gaussian 03. For the neutral state, the MM and QM/MM optimized coordinate sets states give a rmsd of 1.311 (hydrogens excluded). We did not explore the existence of other possible minima. The existence of other local minima close in energy is unlikely because of the circular and very rigid structure of kB1. Only its side chains have some geometric freedom.

Amino acids are referred by the conventional one-letter code hereafter. The protonation state of the two charged amino acids (E3 and R24) was checked using propKa[48,49] (experimental conditions, pH = 6.1).

C. Decomposition into Electrostatic and Geometric Components. In our implementation, the QM wave function is polarized by the electric field created by MM point charges, which is referred to as electrostatic embedding (EE) hereafter. EE can be switched off by setting up all MM point charges to zero: corresponding values are noted $AEA^{\star}$. The two EE-free values, for kB1 and L,L-cystine, serve as calculus intermediates to decompose ΔAEA (eq 2) into geometric and electrostatic contributions. Such that, one can write

$$\begin{aligned}\Delta AEA &= AEA_{\mathrm{kB1}} - AEA_{\mathrm{L,L-cys}} \\ &= \underbrace{(AEA_{\mathrm{kB1}} - AEA^{\star}_{\mathrm{kB1}} - AEA_{\mathrm{L,L-cys}} + AEA^{\star}_{\mathrm{L,L-cys}})}_{\mathrm{elec}} + \\ &\quad \underbrace{AEA^{\star}_{\mathrm{kB1}} - AEA^{\star}_{\mathrm{L,L-cys}}}_{\mathrm{geom}} \\ &= \Delta AEA^{\mathrm{elec}} + \Delta AEA^{\mathrm{geom}}\end{aligned}$$

Rather intuitively, the mechanical constraint exerted by the protein on a cystine fragment corresponds to the difference between kB1 and the linear L,L-cystine, as all MM point charges are turned off. A residue-by-residue analysis of individual side-chain contributions to ΔAEA^{elec} is lead with exactly the same methodology. In contrast with the aforementioned global procedure, backbone point charges are not switched off to avoid the creation of an artificial dipole.[57]

III. Results and Discussion

Disulfide numerotation requires an arbitrary choice because of the circular structure of kB1. We followed the convention of Craik and co-workers, as indicated on Figure 2, with three linkages Cys(1−15), Cys(5−17), and Cys(10−22), where Cys denotes cystine. For the sake of conciseness, they are from now on single-number labeled (respectively SS1, SS2, and SS3) in that order.

A. Respective Reactivities for the One-Electron Addition. First adiabatic EA are reported in Table 1 for each disulfide bridge as well as geometric parameters for neutral and radical anionic forms. First of all, none of the three

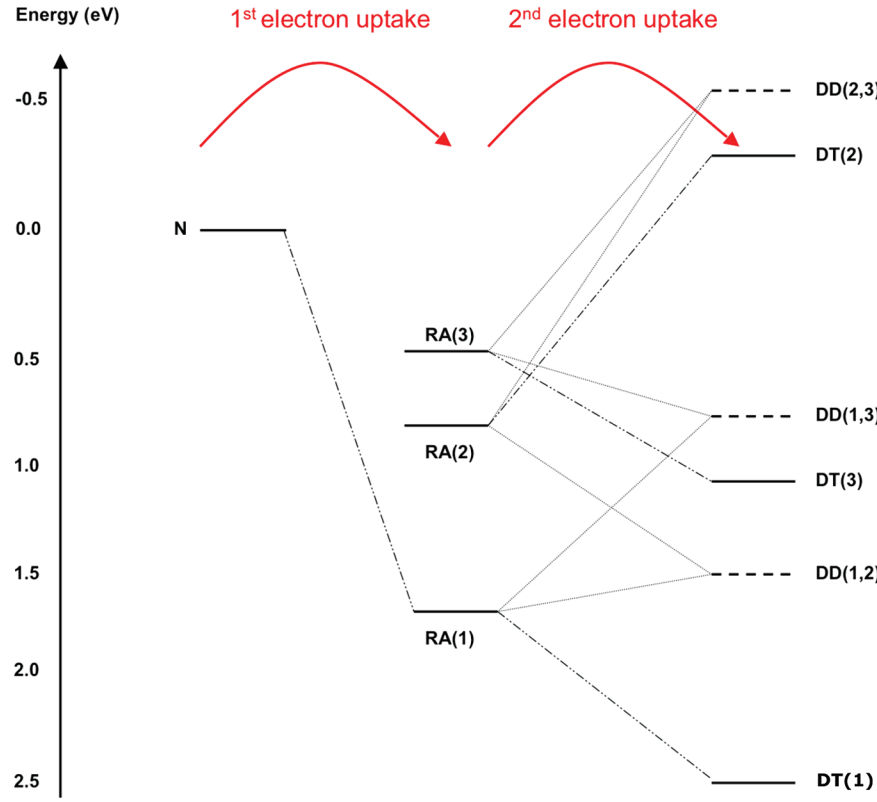


Figure 4. Relative energies (in eV) of kB1 products resulting from successive electron addition(s). RA denotes a (disulfide) radical anion, with the label of linked cysteines. DT (in solid line) refers to nonradical dianions (dithiolates) and DD (dashed lines) diradical dianions with two 2S-3e bonds. For the latter, singlet and triplet states have approximatively the same energies and thus do not appear separately. Corresponding numerical values are reported in Table 2. The neutral system (N) is taken as a reference of energy.

disulfide linkages is dissociated upon electron addition, which has to be noticed as a cleavage of the weak 2S-3e interaction can be observed in a highly dissymmetric environment.[50,51] Moreover, Mulliken spin densities (reported in Supporting Information, Table 1) are almost equally distributed on each sulfur center.

Values of ΔAEAs of +0.52, −0.38, and −0.55 eV are computed respectively for SS1, SS2, and SS3. SS1 is the most reactive toward electron uptake. One can note that Craik et al. proved experimentally that this linkage also exhibits the highest reactivity toward reducing alkylation.[33] We[24−26] and others[5] conjectured a possible analogy between disulfide electron affinity and redox potential. In contrast, the two other disulfide linkages are significantly less prone to capture an electron. How does the proteinic environment tune disulfide electron affinity, which is either increased or decreased with respect to L,L-cystine? To answer this question, ΔAEA are decomposed into electrostatic and geometrical contributions. Their values, gathered in Table 1, indicate that both effects importantly impact on AEA. They are examined in the next two subsections.

B. Residue-by-Residue Decomposition of the Electrostatic Component. The electrostatic modulation from the highly dissymmetric distribution of charge of the protein is an important factor orientating the inner competition for an electron uptake. It is intuitive that the presence of some charged residues in the vicinity of a neutral disulfide is decisive, as ascertained and quantified by previous studies. SS1 is indeed spatially the closest to an arginine, the 24th residue (R24), the sole positively charged residue of kB1 (Figure 2). Yet, its contribution may be counterbalanced by other residues (notably E3, the sole negatively charged one of kB1). Therefore, we performed a more systematic residue-by-residue analysis.

The individual side-chain contributions ΔAEA_l were computed, according to the procedure described in Subsection 2.3. They are monitored in Figure 5, as a function of a *m*-th residue whose side-chain electrostatic contribution is switched off. For comparative purposes, in the intermediate situation where all point charges of the side chain are turned off, but the backbone still polarizes the QM wave function, $AEA_l^{backbone}$, are rather similar (0.74, 0.98, and 0.71 eV, respectively). ΔAEA_l are reported in Table 3 of the Supporting Information as well as distances between disulfide barycenters and C_α positions of each constituting amino acid of kB1 on its optimized geometry. The following comments can be made:

1. As expected, R24 strongly enhances AEA, by 3.11, 1.29, and 1.05 eV for SS1, SS3, and SS2. These increments follow the distances between its C_α, and SS barycenters are 4.83, 8.03, and 10.57 Å, respectively.
2. Conversely, E3 disfavors an electron uptake by 1.76, 1.34, and 2.30 eV, respectively, for SS1, SS2, and SS3, distant by 7.11, 8.47, and 5.34 Å.
3. The decomposition picks out a third *neutral* residue, namely S18, with ΔAEA_l of 0.07, −0.35, and −0.16

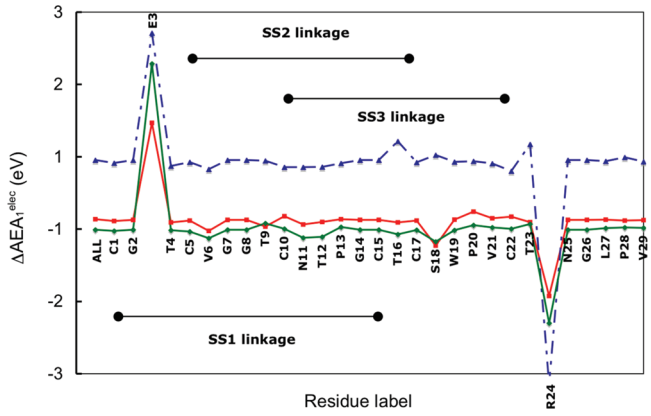


Figure 5. Variations of ΔAEA_1 for each of the three disulfide linkages of kB1, as a function of the residue number whose side chain partial charges are turned off. Values for the most reactive linkage SS1 correspond to the dashed line with (blue) triangles, with SS2 and SS3, respectively, correspond to the (red) square dots and (green) circles. In all cases, the two charged amino acids (E3, R24) exhibit important contributions, whose amplitude depends on their distance to the disulfide barycenter. The remaining of constituting residues of kB1 forms an apolar baseline.

for SS1, SS2, and SS3 linkages. This is most likely related to recent observations that serine, one of the most polar amino acid,[52,53] plays a specific role in tuning the redox potential of -Ser-Cys-Cys-Ser- (SCCS) motifs.[54]

4. In contrast, most of the remaining amino acids in the sequence of kB1 (glycine G, alanine A, valine V, leucine L, isoleucine I, proline P,... usually classified as neutral apolar) form an apolar baseline. They do not significantly impact on EA (variations lower than 0.04 eV in absolute values).

These results draw a simple conclusion, as do Coulomb laws: the closer the residue and the higher its polarity, the stronger its impact on electron affinity. Yet, this should not blur that even nonpolar residues also impact EA, more indirectly, by defining the secondary and tertiary structures. In turn, they create the backbone electrostatic field but also impose a mechanical constraint. The decomposition of ΔAEA_1 clearly denotes the importance of the geometrical effects that are analyzed in the next Subsection.

C. Geometrical Resistance to One-Electron Uptakes. The geometrical contribution ΔAEA^{geom} are quantified with respect to the linear L,L-cystine, for which no steric hindrance exists, and values for each disulfide are reported in the last column of Table 1. Generally speaking, its sign can be either

• positive if the disulfide elongation induced by the one-electron uptake is associated with a release of conformational strain, hence energetically favoring the anionic form. This is often the case of sequentially closed cysteines, like Trx.[24]

• or negative when the drastic disulfide lengthening is geometrically disfavored — for instance in a designed hairpin, with more separated cysteinyl residues.[24]

The negative signs computed for kB1 characterize an energetic penalty that systematically disfavors the anionic form. More quantitatively, SS1 differs from the two others disulfides solely from a geometrical point of view, with a purely mechanical energetic penalty on ΔAEA roughly halved (−8.3 vs −15.9 and −18.2 kcal/mol, values reported in eV in Table 1). This suggests that, as the compact Mobius structure is enforced, the SS1 elongation induces comparatively less defavorable structural changes.[58] The higher malleability of this linkage is further examplified by the variation of dihedral angle τ(C-S-S-C) by 37 degrees. rmsd provide a more global measure of the geometrical reorganization imposed by a disulfide lengthening: values are lower for SS2 and SS3 than for SS1 (respectively 0.23, 0.28, and 0.87 — values in Table 1). This lies in perfect agreement with experimental studies: Craik and co-workers proved that the Ala(1−15) mutant of kB1 conserves a very similar structure to the wild-type protein[32] and came to the conclusion that SS2 and SS3 define the structure of cyclotides, while SS1 is solely responsible for reactivity properties.[32]

D. Second Electron Uptake: A Competition between Dithiolates and Diradical Anions. We now discuss the addition of a second electron, with the competitive formation of a dithiolate or of a second disulfide radical anion (DT vs DD). Electronic energies of all nine possible dianions (three closed-shell dithiolates and six open-shell diradicals, either singlet or triplet) were computed to identify the most stable product. Data are reported in Table 2, and energy levels for neutral, anionic, and dianionic species are displayed on Figure 4. This diagram indicates, with no ambiguity, that the formation of the dithiolate SS1 is strongly favored over other possible adducts, with a spread of ca. 3 eV. The next paragraph explains how the proteinic environment of kB1 induces this orientation.[59]

The two main reasons why this contribution is neglected in this study are as follows: (i) a propKa calculation on the QM/MM optimized structure of DT1 indicating that no proton transfer occurs between the dithiolate DT1 and the R24 residue (whose pK_a is 11.24, to be compared to the reference pK_a of 12.50) and (ii) a large distance (5.60 Å) between the barycenter of S8—S167 and the barycenter of the N—N segment of R24 extremity. Also, whenever existing, such a charge transfer will in the first place stabilize DT1, the lowest energy structure (and the most prone to charge transfer).

Our results clearly show that the ease of reorganization of the protein upon disulfide elongation (by ca. 0.7 Å as a radical anion is formed, or by at least 2 Å for the formation of a dithiolate) is a decisive factor for the stabilization of a dianion. We first limit the discussion to dithiolates. DT1 is the most stable entity, by 1.38 and 2.72 eV over DT3 and DT2: the larger the distance between the two negatively charged sulfurs, the lower the energy. (Intersulfur distances, reported in Table 2 are respectively 6.68, 5.77, and 4.70 Å.) The local rigidity of the structural SS2 linkage prevents a spatial separation needed to stabilize the dianion. In contrast, a close inspection of the optimized geometry of the SS1 dithiolate (Figure 6c), compared to the structures of the neutral and anionic species (Figure 6a,b), reveals a different orientation of the R24 side chain. Its positively charged end $-CH(NH_2)_2^+$ points in the direction of the cleaved 1−15 disulfide and helps to stabilize one of the thiolates (C15). This motion is associated with the formation of a new

Table 2. Geometries, RMSD and Relative Energies (in eV) of Dianionic Forms of kB1 - Dithiolates (DT) or Diradical Dianions (DD)[a]

compounds			structure				
	2S+1	linkage(s)	d(S−S)	∠(S-S-C)	τ(C-S-S-C)	rmsd	energy ΔE
			Dithiolates				
DT1	1	SS1	6.68	70.9, 120.5	146.8	1.12	2.47
DT2	1	SS2	4.70	76.2, 83.1	78.2	0.57	−0.25
DT3	1	SS3	5.77	39.6, 109.3	114.2	1.44	1.09
			Diradical Dianions				
DD(1,2)	1	SS1	2.76	92.1, 99.3	128.8	0.83	0.88
		SS2	2.76	91.2, 96.4	68.8		
	3	SS1	2.76	92.1, 99.3	128.8	0.83	0.88
		SS2	2.76	91.2, 96.4	68.8		
DD(1,3)	1	SS1	4.97	81.7, 113.8	135.0	1.05	0.40
		SS3	2.75	98.1, 93.6	120.6		
	3	SS1	5.97	94.1, 106.0	120.2	1.51	0.40
		SS3	2.76	90.6, 90.6	125.3		
DD(2,3)	1	SS2	2.76	94.9, 92.4	62.3	0.64	−0.50
		SS3	2.71	50.3, 91.4	118.3		
	3	SS2	2.76	94.9, 92.4	62.3	0.64	−0.51
		SS3	2.72	94.2, 91.4	118.4		

[a] Mulliken spin densities are given in Table 1, Supporting Information. Relative energies ΔE are calculated with the neutral (N) compound taken as a reference — cf. Figure 4.

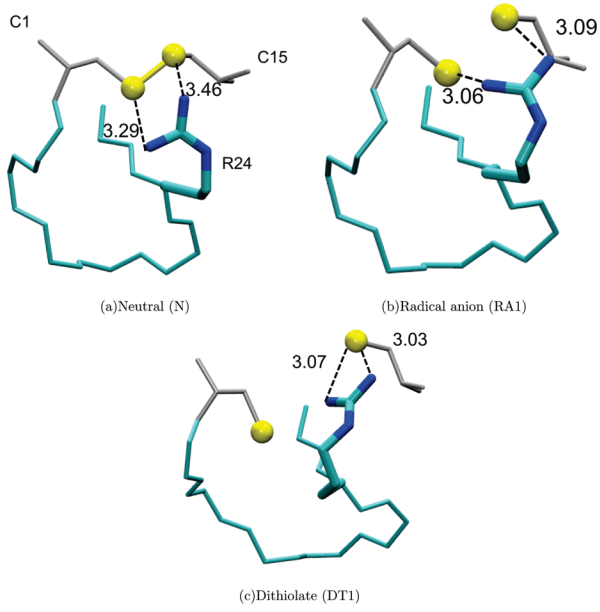


Figure 6. Partial view of kB1 optimized structures, centered on the SS1 linkage. The backbone is displayed with (green) light sticks, and arginine 24 side chain (R24), which plays a crucial role in tuning the one- and two-electron uptake, in bolder sticks. We report for each structure the two lower distances (in Å) between nitrogens of R24 and sulfurs of C1 and C15. A dissymmetry appears as the disulfide bond is disrupted (dithiolate DT1), and the side chain of R24 stabilizes the C15 sulfur thiolate.

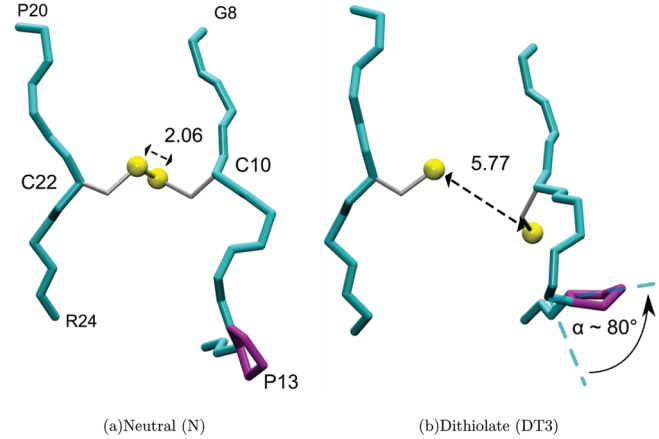


Figure 7. Optimized structures of kB1, in the neutral (left side) and dianionic (right side) forms, centered on the SS3 linkage. The backbone is depicted with (green) sticks, and the side chain of proline P13 with purple sticks, while the two sulfur atoms are displayed with (yellow) balls. Distances are reported in Å. One notes a large amplitude motion of the P13-G14 β-turn, with a characteristic angle α of ca. 80 degrees.

hydrogen bond network in the vicinity of R24. Both effects counterbalance the repulsion of the two negatively charged sulfur atoms. In between, DT3 is also stabilized by a flip of the P13-G14 β-turn upon the 3.71 Å elongation of the initially covalent SS — represented in Figure 7. This large amplitude motion (rmsd value of 1.51, characterized by an α angle of ca. 80 degrees) induces a flip of the -CH_2-S^- side chain of C10. Its sulfur atom rotates to point oppositely to the other C22 sulfur, whose position remains almost unchanged.

The most stable of the three diradical dianions, DD(1,2), could have been predicted from the reactivity order for the first electron uptake (AEA_1, cf. Figure 4). Most likely, the presence of E3 governs the energetic positions of DD(1,2) vs DD(1,3).

The latter compound examplifies an interesting structural outcome. The electron addition on the SS3 linkage induces a cleavage of the 2S-3e bond of the SS1 radical anion, with a distance passing from 2.75 to 4.97 Å. R24, initially equidistant to each sulfur of the SS1 linkage (Figure 6b), stabilizes one of the thiolates (C1) — the situation is close to Figure 6c. This evolution can be related to the charge-assisted electron capture dissocation of disulfide, studied by both experimental[55] and theoretical means.[50]

One should keep in mind that the formation of a dithiolate, even energetically favored, would probably not be observed if crystals of kB1 were irradiated, because the geometrical relaxation is hindered/prevented in a crystallic structure (packing effect). Our calculations provide a complementary

view on the competitive formation of one- and two-electron addition adducts.

IV. Concluding Remarks

In this work, we examplified on the prototypic cyclotide, kalata B1, how a proteinic environment can dramatically tune the one- and two-electron reactivity of disulfide linkages. The factors governing the inner competition for the valence-attachment or the formation of a dithiolate vs a second disulfide radical anion, concomitant to a partial unfolding, are traced back. *Both* the electrostatic field (largely dominated by the respective positions of R24, E3, and, in a lesser extent, S18) and the mechanical constraint intermingle to increase the electron affinity of one disulfide of the cystine knot motif for the one-electron addition. This decomposition may provide useful information for guiding experimental works aiming at understanding and ultimately tuning in a controlled way disulfide reactivity (systematic scanning mutagenesis on cyclotides).[56]

Comparison is made with experimental studies that also provide strong evidence for the reactive role of this linkage, in terms of redox potential. It is quite remarkable that the same factors seem to govern electron affinity or redox potential. This nascent similarity of behaviors deserves more systematic studies, in order to draw a parallel that might lead to a more unified view of disulfide reactivity.

Acknowledgment. This work was supported by computer resources of the University of Nancy I, France. The authors are grateful to Dr. Nicolas Ferré (Université Aix-Marseille I, France) for providing a local version of the link atom scheme. A.D.L. and X.A. also acknowledge financial support from the Jean Barriol Institute (FR CNRS 2843). One of us (P.F.L.) thanks the Australian Research Council (ARC) for a postdoctoral fellowship.

Supporting Information Available: Mulliken spin densities of the nine radical structures, structures and electron affinities of neutral, radical anions and dithiolates described with a single -CH_2-S-S-CH_2- QM part, and numerical data corresponding to Figure 5. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Pauling, L. *J. Am. Chem. Soc.* **1931**, *53*, 1367–1400.

(2) Gill, P. M. W.; Radom, L. *J. Am. Chem. Soc.* **1987**, *110*, 4931–4941.

(3) Asmus, K. D. *Acc. Chem. Res.* **1979**, *12*, 436–442.

(4) Tung, T.-L.; John Stone, A. *Can. J. Chem.* **1975**, *53*, 3153–3157.

(5) Houée-Levin, C.; Bergès, J. *Radiat. Phys. Chem.* **2008**, *77*, 1286–1289.

(6) Fang, X.; Wu, J.; Wei, G.; Schuchmann, H. P.; von Sonntag, C. *Int. J. Radiat. Biol.* **1995**, *68*, 459–466.

(7) Lawrence, C. C.; Bennati, M.; Obias, H. V.; Bar, G.; Griffin, R. G.; Stubbe, J. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 8979–8984.

(8) Johnson, D. L.; Polyak, S. W.; Wallace, J. C.; Martin, L. L. *Peptide Sci.* **2003**, *10*, 495–500.

(9) Chen, X.; Zhang, L.; Wang, Z.; Li, J.; Wang, W.; Bu, Y. *J. Phys. Chem. B* **2008**, *112*, 14302–14311.

(10) Lao, Y.-T.; Abu-Irhayem, E.; Kraatz, H.-B. *Chem.–Eur. J.* **2005**, *11*, 5186–5194.

(11) Glese, B.; Graber, M.; Cordes, M. *Curr. Opin. Chem. Biol.* 2008. asap.

(12) Gauduel, Y.; Gelabert, H.; Guilloud, F. *J. Am. Chem. Soc.* **2000**, *122* (21), 5082–5091.

(13) Gauduel, Y.; Marignier, J. L.; Belloni, J.; Gelabert, H. *J. Phys. Chem. A* **1997**, *101* (48), 8979–8986.

(14) Gauduel, Y.; Launay, T.; Hallou, A. *J. Phys. Chem. A* **2002**, *106*, 1727–1732.

(15) Antonello, S.; Benassi, R.; Gavioli, G.; Taddei, F.; Maran, F. *J. Am. Chem. Soc.* **2002**, *124*, 7529–7538.

(16) Antonello, S.; Daasbjerg, K.; Jensen, H.; Taddei, F.; Maran, F. *J. Am. Chem. Soc.* **2003**, *125*, 14905–14916.

(17) Wenska, G.; Filipiak, P.; Asmus, K.; Bobrowski, K.; Koput, J.; Marciniak, B. *J. Phys. Chem. B* **2008**, *112*, 10045–10053.

(18) Ya. Melńikov, M.; Weinstein, J. A. *High Energy Chem.* **2008**, *42*, 329–331.

(19) Carles, S.; Lecomte, F.; Schermann, J.-P.; Desfrancois, C.; Xu, S.; Milles, J. M.; Bowen, K. H.; Bergès, J.; Houée-Levin, C. *J. Phys. Chem. A* **2001**, *105*, 5622–5626.

(20) Weik, M.; Bergès, J.; Raves, M. L.; Gros, P.; McSweeney, S.; Silman, I.; Sussman, J. L.; Houée-Levin, C.; Ravelli, R. B. G. *J. Synchroton Radiat.* **2002**, *9*, 342–346.

(21) Weik, M.; Ravelli, R. B.; Silman, I.; Sussman, J. L.; Gros, P.; Kroon, J. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 623–628.

(22) Fourre, I.; Silvi, B. *Heteroat. Chem.* **2007**, *18*, 135–160.

(23) Dumont, E.; Loos, P.-F.; Assfeld, X. *Chem. Phys. Lett.* **2008**, *458*, 276–280.

(24) Dumont, E.; Loos, P.-F.; Assfeld, X. *J. Phys. Chem. B* **2008**, *112*, 13661–13669.

(25) Dumont, E.; Loos, P. F.; Laurent, A. D.; Assfeld, X. *Int. J. Quantum Chem.* 2009. in press.

(26) Dumont, E.; Loos, P.-F.; Laurent, A. D.; Assfeld, X. *J. Chem. Theory Comput.* **2008**, *4*, 1171–1173.

(27) Sawicka, A.; Berdys-Kochanska, J.; Skurski, P.; Simons, J. *Int. J. Quantum Chem.* **2005**, *102*, 838–846.

(28) Rickard, G. A.; Bergès, J.; Houée-Levin, C.; Rauk, A. *J. Phys. Chem. B* **2008**, *112*, 5774–5787.

(29) Ireland, D. C.; Wang, C. K.; Wilson, J. A.; Gustafson, K. R.; Craik, D. J. *Peptide Sci.* **2007**, *90*, 51–60.

(30) Shenkarev, Z. O.; Nadezhdin, K. D.; Sobol, V. A.; Sobol, A. G.; Skjeldal, L.; Arseniev, A. S. *FEBS* **2006**, *273*, 2658–2672.

(31) Colgrave, M. L.; Craik, D. J. *Biochemistry* **2004**, *43*, 5965–5975.

(32) Daly, N. L.; Clark, R. J.; Craik, D. J. *J. Biol. Chem.* **2003**, *8*, 6314–6322.

(33) Goransson, U.; Craik, D. J. *J. Biol. Chem.* **2003**, *48*, 48188–48196.

(34) Clark, R. J.; Daly, N. L.; Craik, D. J. *Biochem. J.* **2006**, *394*, 85–93.

(35) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(36) Braida, B.; Hiberty, P. C.; Savin, A. *J. Phys. Chem. A* **1998**, *102*, 7872–7877.

(37) Humbel, S.; Demachy, I.; Hiberty, P. C. *Chem. Phys. Lett.* **1995**, *247*, 126–134.

(38) Rienstra-Kiracofe, J. C.; Tschumper, G. S.; Schaefer, H. F., III.; Nand, S.; Ellison, G. B. *Chem. Rev.* **2002**, *102*, 231–282.

(39) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.

(40) Dapprich, S.; Komárino, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (THEOCHEM)* **1999**, *461*, 1–21.

(41) MacKerel, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B., III.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(42) Brooks, B. R.; Bruccoleri, R. E.; Olafson, D. J.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(43) MacKerell, A. D., Jr.; Brooks, C. L., III.; Nilsson, L.; Roux, B.; Won, Y.; Karplus M. John Wiley & Sons: Chichester, 1998; Vol. 1of The Encyclopedia of Computational Chemistry; p 271.

(44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision B.05*; Gaussian, Inc.: Wallingford, CT, 2004.

(45) Ponder, J. W. *Tinker, version 4.2*; Washington University: St. Louis, MO, 2004.

(46) Kabsch, W. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1978**, *A34*, 827–828.

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

(48) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(49) Bas, D. C.; Rodgers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765–783.

(50) Sawicka, A.; Skurski, P.; Hudgins, R. R.; Simons, J. *J. Phys. Chem. B* **2004**, *107*, 13505–13511.

(51) Anusiewicz, I.; Berdys-Kochanska, J.; Simons, J. *J. Phys. Chem. A* **2005**, *109*, 5801–5813.

(52) Grantham, R. *Science* **1974**, *185*, 862–864.

(53) Zimmerman, J. M.; Eliomi, N.; Simha, R. *J. Theor. Biol.* **1968**, *21*, 170–201.

(54) Gromer, S.; Johansson, L.; Bauer, H.; Arscott, L. D.; Rauch, S.; Ballou, D. P.; Williams, C. H., Jr.; Heiner Schirmer, R.; Arner, E. S. J. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9533–9538.

(55) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.

(56) Simonsen, S. M.; Sando, L.; Rosengren, K. J.; Wang, C. K.; Colgrave, M. L.; Daly, N. L.; Craik, D. J. *J. Biol. Chem.* **2008**, *283*, 9805–9813.

(57) If the point charges of an *i* residue were switched to zero, an artificial dipole would be created by the backbone nitrogen of the *i*-1 residue and the carbon of the $i + 1$ residue — corresponding to charges of -0.21 and $+0.21$ a.u., at a distance of ca. 5.5 Å.

(58) NMR 1H chemical shifts of kB1 also indicate '*a less structured local conformation*'.[32]

(59) One of the main drawbacks of QM/MM methods is to break artificially the possibility of a charge transfer. Two types of charge transfer can be considered, through the frontier bonds between the QM and the MM parts and through space. The excess electron is placed in the well localized $\sigma^\star$ (SS) orbital, and hence the charge transfer through the frontier bond is expected to be very small. (A MP2/6-31+G** calculation on diethyldisulfide shows that the terminal methyl groups have their Mulliken charges changed by less than 0.004e upon electron capture). For the through space charge transfer, several processes can occur. The two main reasons why this contribution is neglected in this study are as follows: (i) a propKa[48,49] calculation on the QM/MM optimized structure of DT1 indicating that no proton transfer occurs between the dithiolate DT1 and the R24 residue (whose pKa is 11.24, to be compared to the reference pKa of 12.50) and (ii) a large distance (5.60 Å) between the barycenter of S_8—S_{167} and the barycenter of the N—N segment of R24 extremity. Finally, whenever existing, such a charge transfer will in the first place stabilize DT1, the lowest energy structure (and the most prone to charge transfer), and thus not affect our conclusions.

CT900093H

JCTC Journal of Chemical Theory and Computation

Interaction of Benzene with Transition Metal Cations: Theoretical Study of Structures, Energies, and IR Spectra

Hai-Bo Yi,*,† Han Myoung Lee,*,‡ and Kwang S. Kim*,§

College of Chemistry and Chemical Engineering, Hunan University, Changsha, Hunan 410082, China, Department of Chemistry and Center for Basic Sciences, Pohang University of Science and Technology, Pohang 790-784, Korea, and Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

Received March 31, 2009

Abstract: The cation$-\pi$ interactions have been intensively studied. Nevertheless, the interactions of π systems with heavy transition metals and their accurate conformations are not well understood. Here, we theoretically investigate the structures and binding characteristics of transition metal (TM) cations including novel metal cations (TM^{n+} = Cu^+, Ag^+, Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+}) interacting with benzene (Bz). For comparison, the alkali metal complex of Na^+-Bz is also included. We employ density functional theory (DFT) and high levels of ab initio theory including Møller−Plesset second-order perturbation (MP2) theory, quadratic CI method with single and double substitutions (QCISD), and the coupled cluster theory with single, double, and perturbative triple excitations (CCSD(T)). Each of the transition metal complexes of benzene exhibits intriguing binding characteristics, different from the typical cation$-\pi$ interactions between alkali metal cations and aromatic rings. The complexes of Na^+, Cu^+, and Ag^+ favor the conformation of C_{6v} symmetry with the cation above the benzene centroid (π_{cen}). The formation of these complexes is attributed to the electrostatic interaction, while the magnitude of charge transfer has little correlation with the total interaction energy. Because of the $TM^{n+}\leftarrow\pi$ donation, cations Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} prefer the off-center π conformation (π_{off}) or the π coordination to a C atom of the benzene. Although the electrostatic interaction is still important, the $TM\leftarrow\pi$ donation effect is responsible for the binding site. The $TM^{n+}-Bz$ complexes give some characteristic IR peaks. The complexes of Na^+, Cu^+, and Ag^+ give two IR active modes between 800 and 1000 cm^{-1},which are inactive in the pure benzene. The complexes of Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} give characteristic peaks for the ring distortion, C−C stretching, and C−H stretching modes as well as significant red-shifts in the CH out-of-plane bending.

I. Introduction

Cation$-\pi$ interactions have been characterized in a wide range of contexts,[1−4] due to the importance in diverse fields of chemistry,[5] biology,[6] and nanotechnology.[7] A number of studies have been reported on the binding of alkali metal cations or organic cations with ethylene, acetylene, benzene, or other π systems. These interaction forces have been utilized to design ionophores and receptors.[8] Benzene (Bz) is a good prototype aromatic compound and serves as a model for the π systems. Alkali metal cations prefer the formation of π complexes of C_{6v} symmetry.[9] As transition metal arene complexes that are key intermediates in aromatic C−H bond activation display multifaceted coordination chemistry,[11] transition metal complexes with aromatic compounds have been widely investi-

* Corresponding author e-mail: hbyi@hnu.cn (H.-B.Y.); abcd0lhm@postech.ac.kr (H.M.L.); kim@postech.ac.kr (K.S.K.).

† Hunan University.

‡ Department of Chemistry and Center for Basic Sciences, Pohang University of Science and Technology.

§ Department of Chemistry, Pohang University of Science and Technology.

10.1021/ct900154x CCC: $40.75
Published on Web 05/11/2009

gated,[10] and the interactions of novel metals (Cu^+, Ag^+, and Au^+) with Bz have been reported.[12]

Recently, the interest in gas-phase reactions for metal dications has grown[13−15] due to advances in experimental techniques such as electrospray ionization and electron impact double ionization, which permit the generation of dication in the gas phase. Using NMR spectroscopy, Johansson et al. detected a Pt^{2+}−benzene complex, [(diimine)Pt(η^2-C_6H_6)CH_3]$^+$, a precursor to arene C−H oxidative addition.[15] Templeton and co-workers gave further structural characterization for the Pt^{2+} η^2-benzene adduct.[16] In this regard, a further detailed investigation of binding features of the TM^{n+}−Bz complexes would be of importance.

It is known that the interaction of alkali−metal cations with Bz is mainly governed by the electrostatic and induction interactions.[1,5,17] In the interactions of olefinic, aromatic, and heteroaromatic π systems with alkali metal ions[5] and also in the $(C_2H_4−TM)^+$ complexes,[18,19] the electrostatic interaction also plays an important role. However, molecular dications formed by attachment of a TM^{2+} dication to a neutral base often show significant bonding features. Thus, it is important to understand the role of ionic/covalent bonding in the formation of a $TM^{2+}−\pi$ cluster.

In this study, we investigate the bindings of Pd^{2+}, Pt^{2+}, and Hg^{2+} with Bz using ab initio theory and density functional theory, and we also report the bindings of Na^+, Cu^+, Ag^+, and Au^+ with Bz for comparison. It is vital to understand these interactions for the design and development of the receptors and sensors for the heavy transition metal recognition as well as the hazardous biological problems of the heavy transition metal intercalation between DNA stacks, which have been hot topics in molecular/biomolecular recognition study of heavy metals.

The formation of TM^{n+}−Bz complexes is often associated with charge transfer from TM^{n+} to Bz. The charge transfer affects the binding feature and structural distortion of π moieties. Both geometrical and electronical (i.e., charge-transfer/polarization) changes due to the complexation of TM^{n+} with Bz result in significant changes in IR spectra. Thus, to facilitate the experiments for the structural information of $TM^{n+}−\pi$ complexes, we also compare the differences in IR spectra between different transition metal complexes.

II. Methods

Many possible structures of TM^{n+}−Bz complexes were optimized using DFT calculations [Becke three parameters with the Lee−Yang−Parr functional (B3LYP) and the Perdew, Burker, and Ernzerhof functional (PBE)] and ab initio calculations [Möller Plesset second-order perturbation theory (MP2), and quadratic CI method including single and double substitutions (QCISD)].[20] The basis set for benzene was used with the aug-cc-pVDZ (aVDZ) basis set. The pseudopotentials of transition metals (Cu, Ag, Au, Pd, Pt, and Hg) were used with the Stuttgart RSC 1997 effective core potential (ECP).[21] For Au, Pd, Pt, and Hg, the relativistic effective core potentials (RECP) developed by the Stuttgart group were used in conjunction with the basis set to describe the metal valence electrons. For Na, the cc-pCVDZ basis set was used, and a single f function was added for Cu, Ag, and Au.[22] The binding energies were further calculated using the coupled cluster theory with single, double, and perturbative triple excitations (CCSD(T)) employing the aVDZ basis set and the MP2 theory using the aug-cc-pVTZ (aVTZ) basis set (for benzene) on the MP2/aVDZ optimized geometries. In this aVTZ case, the cc-pCVTZ basis set was used for Na, and a set of two f and one g polarization functions were added for transition metals (Cu, Ag, Au, Pd, Pt, and Hg), as suggested by Martin and Sundermann.[23] The basis set superposition error (BSSE) correction was taken into account. The complete basis set (CBS) limit values for the MP2 binding energies were evaluated on the basis of the extrapolation method to exploit that the electron correlation error is proportional to N^{-3} for the aug-cc-pV*N*Z basis set.[24] Given that the difference in binding energy between MP2 and CCSD(T) for the same basis set does not change significantly with increasing basis set size, the CCSD(T)/CBS binding energies were evaluated from the MP2/CBS ones by applying the difference between CCSD(T) and MP2 binding energies for the aug-cc-pVDZ basis set.[24b] QCISD calculations were carried out using the 6-31G** basis set, with the relativistic effective core potentials (RECP) and the corresponding basis set for transition metals. All calculations in this work were carried out using a suite of Gaussian 03 programs.[25] Molecular orbital analysis was done on the basis of the MP2 calculations by using the POSMOL package.[26]

III. Results

A. TM Cations and Benzene. Novel metals (Cu^+, Ag^+, and Au^+) as well as alkali metal Na^+ have singlet ground states, and their complexes also have singlet ground states. Pd^{2+} and Pt^{2+} have triplet ground states of the d^8 type with electronic configurations of $4d^8$ and $4f^{14}5d^8$, respectively. The Pd^{2+}−Bz and Pt^{2+}−Bz complexes may exist in either the triplet or singlet state. Hg^{2+} is of the d^{10} type with an electronic configuration of $4f^{14}5d^{10}$; thus, the Hg^{2+} complexes are mainly closed shell systems with the singlet ground state. The "5d" atomic orbitals (AOs) of Hg^{2+} are fully occupied, so that the interactions of Hg^{2+} with π ligands would be quite different from those of Pd^{2+} and Pt^{2+}.

For the ionization potentials (IP) of neutral metals, the B3LYP/aVDZ and PBE/aVDZ values are slightly overestimated, while QCISD/aVTZ values are slightly underestimated. The MP2/aVTZ and CCSD(T)/aVTZ values are in good agreement with the experimental values (except for the case of Hg),[27] as listed in Table 1. However, as for the electron affinity (EA), the B3LYP/aVDZ and PBE/aVDZ values are in reasonable agreement with the experimental values,[27] the MP2/aVTZ values are slightly underestimated, and the CCSD(T)/aVTZ values are very close to the experimental values except for the case of Hg. Because of the relativistic effects, the energy of 6s is lowered, and the energy splitting between "5d" and "6s" AOs of Au is much smaller than that between "4d" and "5s" AOs of Ag. As a result, the IP value of Au is much larger than those of Cu and Ag. Similarly, the large IP of Pt or Hg is also attributed to the relativistic effect. Despite that the CCSD(T)/aVTZ well reproduces the experimental

Table 1. Ionization Potential (IP) and Electron Affinities (EA) for TM = Na, Cu, Ag, Au, Pd, Pt, and Hg (in eV at the B3LYP/aVDZ, MP2/aVTZ, and CCSD(T)/aVTZ Levels)[a]

		IP						EA						d−s gap[b]
	conf.	B3LYP	PBE	MP2	CCSD(T)	QCISD	expt.	B3LYP	PBE	MP2	CCSD(T)	QCISD	expt.	MP2
Na	$2p^{10}3s^1$	5.40	5.33	5.07	5.08	4.95	5.14	0.59	0.56	0.21	0.44	0.53	0.55	
Cu	$3d^{10}\ 4s^1$	8.03	8.15	7.59	7.38	7.49	7.73	1.00	1.02	0.81	0.86	1.05	1.24	1.56
Ag	$4d^{10}\ 5s^1$	7.97	8.05	7.47	7.41	7.33	7.57	1.38	1.39	1.00	1.18	1.09	1.30	3.55
Au	$5d^{10}\ 6s^1$	9.33	9.42	9.13	8.93	8.83	9.23	2.21	2.28	2.18	2.07	1.93	2.31	1.89
Pd	$4d^{10}$	8.70	8.80	8.62	8.14	8.03	8.34	0.77	0.92	0.24	0.39	0.32	0.56	2.51
Pt	$5d^9\ 6s^1$	9.25	9.36	9.08	8.80	8.74	9.00	1.39	1.98	1.25	1.91	0.89	2.13	1.01
Hg[c]	$5d^{10}\ 6s^2$	9.70	9.23	9.63	9.49	9.37	10.44	−0.22	−0.20	−0.46	−0.84	−0.45	0	5.17[c]

[a] The experimental IP and EA values are from ref 27. [b] Energy gap between $(n-1)$d and ns orbitals. [c] Inert-pair effect of $6s^2$.

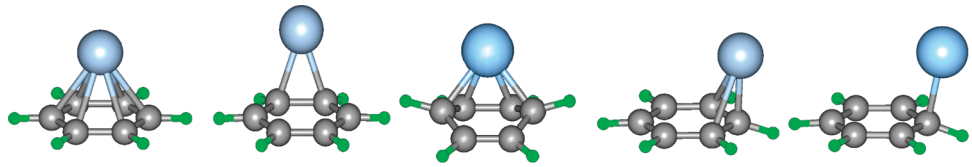


Figure 1. Binding sites of TM^{n+} for Bz.

IP and EA of various metals in a very consistent manner, it gives significant deviation in the IP and EA only for Hg possibly due to the basis set insufficiency. Nevertheless, the structure and binding energy for the Hg^{2+}−benzene complex would still be reliable because the binding energy reflects the cancellation effect of errors and the basis set would not be insufficient for Hg^{2+}, which has two electrons less than Hg (even though it is insufficient for Hg). The s−d energy splitting of Hg is particularly large due to the insignificant spin−orbital coupling because Hg has full 5d and 6s valence shells. In the case of Pd^{2+}/Pt^{2+}, an electron from a singly occupied "d" AO can be promoted to an unoccupied "s" AO, which makes it possible a π←d donation for $Pd^{2+}/Pt^{2+}-\pi$ complexes.

B. TM^{n+}−Benzene Complexes: Conformation. We have investigated several different conformations for the binding of TM^{n+} with Bz using B3LYP/aVDZ and MP2/aVDZ calculations. The conformations for these complexes can be classified into five different binding types [$\pi_{cen}(\eta^6)$, $\pi_{off}(\eta^4)$, $\pi_{off}(\eta^2)$, $\pi_C(\eta^3)$, and $\sigma_C(\eta^1)$], as shown in Figure 1. In the case of π_{cen}, the TM is above the ring centroid, interacting with six carbon atoms, η^6. For π_{off}, the TM is above the center of a carbon bond (η^2) or above two carbon bonds (η^4). In the case of π_C, the TM is above three carbon atoms of Bz (η^3). For σ, the TM is above one carbon atom of Bz (η^1). The binding of Na^+ with Bz is also reported for comparison. For convenience's sake, we put Na in TM in terms of notation. For the binding of TM^{n+} to Bz, the Na^+ cation favors the π_{cen} conformation. The Cu^+ cation can change the position above the whole Bz plane, as can be noted from very small differences between different conformations, and, accordingly, the lowest energy conformation depends on the calculation level of theory. At the MP2/aVDZ level, the π_{cen} structure is slightly more favored. For the Ag^+ cation, B3LYP/aVDZ, PBE/aVDZ, and QCISD/6-31G** favor the π_{off} and π_C structures, but MP2/aVDZ favors the π_{cen} structure. The Au^+ cation favors the π_{off} structure. The Pd^{2+} and Pt^{2+} cations favor the π_{off} conformation for the singlet state and the π_{cen} conformation for the triplet state.

The structures of the TM^{n+}−Bz complexes at the MP2/aVDZ level are given in Figure 2. Upon the complexation (π_{cen}) of the benzene with Na^+, Cu^+, and Ag^+, the C−C bond lengths are slightly increased (by 0.005, 0.017, and 0.012 Å, respectively). In the $\pi_{off}(\eta^2)$ complexes of Au^+, one C−C bond length increases up to 1.461 Å, while in the $\pi_{off}(\eta^4)$ complexes of Pd^{2+} and Pt^{2+}, four C−C bond lengths increase up to 1.466−1.467 and 1.473−1.480 Å, respectively. For the Hg^{2+} complex, the $\pi_C(\eta^3)$ is slightly more favored than $\pi_{off}(\eta^4)$. The π_{off} and π_C conformers have some of the covalent characters (i.e., σ conformers), which are similar to the protonated Bz complex.[28]

The binding energies of TM^{n+}−Bz complexes are in Table 2. The Na^+−Bz system has been investigated intensively at various theoretical levels.[29] The most stable structure of Na^+−Bz is of C_{6v} symmetry, the B3LYP/aVDZ binding energy is 20.4 kcal/mol, and the MP2/CBS value is 21.7 kcal/mol. The CCSD(T)/CBS value (22.1 kcal/mol) is in excellent agreement with the experimental value (22.1 kcal/mol).[30] For the Cu^+−Bz complex, the B3LYP/aVDZ and PBE/aVDZ binding energies of π_{cen}, π_{off}, and π_C structures are nearly same, whereas at the MP2/aVDZ and QCISD/6-31G* levels only the π_{cen} structure is stable. For the Ag^+−Bz complex, the B3LYP/aVDZ, PBE/aVDZ, and QCISD(T)/6-31G* binding energies of the π_{off} and π_C structures are greater than that of π_{cen} structure, whereas at the MP2/aVDZ, MP2/CBS, and CCSD(T)/CBS levels only the π_{cen} structure is stable. For the Au^+−Bz complex, the π_C and π_{off} structures are similar in energy, which are much more stable than the π_{cen} structure. Although the B3LYP/aVDZ, MP2/aVDZ, and CCSD(T)/aVDZ binding energies are underestimated, the PBE/aVDZ and MP2/CBS binding energies are overestimated, and the MP2/aVTZ and CCSD(T)/CBS binding energies are in good agreement with the experimental values. The QCISD/6-31G** binding energy of Na^+−Bz agrees well with the experimental values, while the QCISD/6-31G**

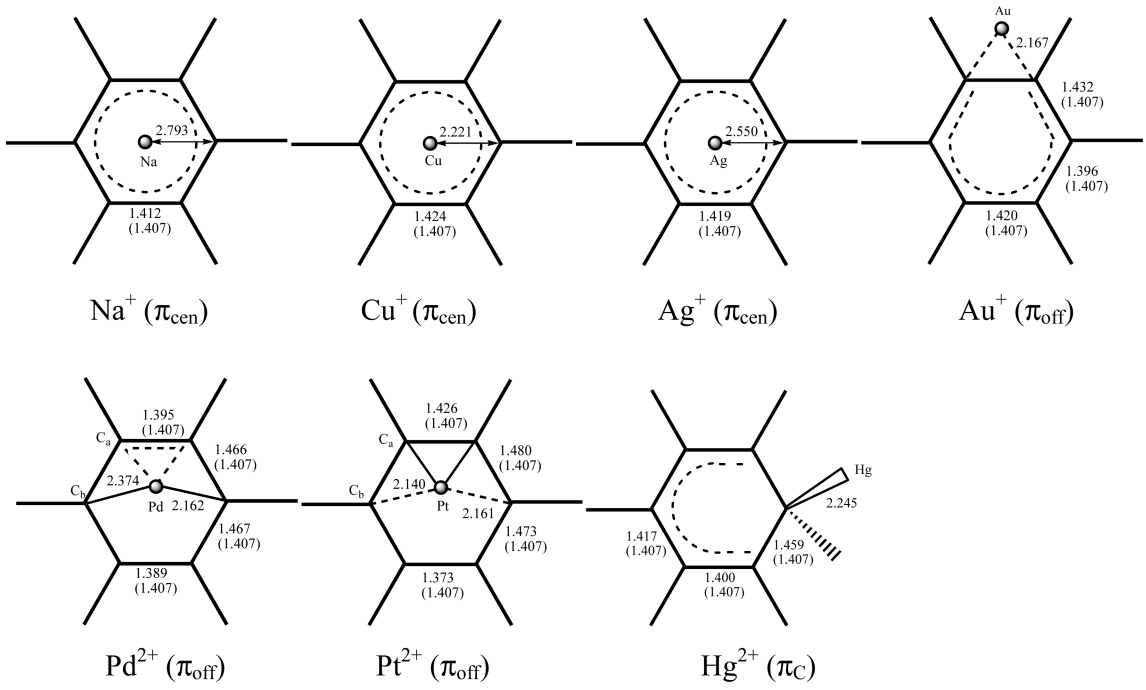


Figure 2. Most stable $TM^{n+}-Bz$ complexes at the MP2/aVDZ level. Bond lengths are in angstroms. The C−C bond length of the uncomplexed Bz is 1.407 Å.

calculations apparently underestimate those of Cu^+-Bz and Au^+-Bz. Pt^{2+} has the singlet ground state due to its bonding feature. However, for $Pd^{2+}-Bz$, the triplet state of the π_{cen} structure is more stable at the B3LYP, PBE, and QCISD levels, while the singlet state of the π_{off} structure is more stable at the MP2/aVDZ, MP2/aVTZ, MP2/CBS, CCSD(T)/aVDZ, and CCSD(T)/CBS levels. Thus, at the present level of theory, the π_{off} structure is more likely, but for clear conclusion, this structure needs to be further investigated at much higher levels of theory in the future.

The vibrational frequencies of Bz were calculated using B3LYP/aVDZ, MP2/aVDZ, and QCISD/6-31G** (Table 3 and Figure 3). At the MP2/aVDZ level, the strong peaks at 678, 1052, 1469, and 3229 cm^{-1} are assigned as an out-of-plane H bending mode, an in-plane H bending mode, a ring distortion mode, and a CH stretching mode, respectively, which correspond to the strong experimental bands observed at 673, 1038, 1469, and 3210 cm^{-1} by Jaeger et al.[33]

C. Binding Characteristics of the TM^{n+} Cation with Benzene. As compared to the nonpolarizable cation Na^+, the binding energy of Cu^+, Ag^+, or Au^+ with Bz is relatively large. However, the formation of $TM^{n+}-Bz$ complexes often shows significant bonding characteristics associated with the charge transfer from TM^{n+} to Bz. The atomic charges were calculated using MP2/aVTZ calculations based on the natural bond orbital (NBO) population analysis (Table 4).

C1. Binding of Na^+, Cu^+, and Ag^+ to Benzene. Cu^+ and Ag^+ have the electron configuration of d^{10}, and so the $TM \leftarrow \pi$ donation in these complexes would not be significant. The experimental IP values of Cu and Ag are 7.73 and 7.57 eV,[27] respectively (7.59 and 7.47 eV at the MP2/aVTZ level, respectively). The experimental IP of Bz is 9.3 eV[35] (9.11 eV at the MP2/aVTZ level), and thus the donation from the π electrons of Bz to the unoccupied s orbital would not be favorable. Based on the NBO charge population, the charge transfer from Cu^+ or Ag^+ to Bz is insignificant (Table 4). Thus, the binding between Bz and $Na^+/Cu^+/Ag^+$ is attributed to the electrostatic and inductive energies, and this binding decreases with increasing distance between the metal cation and the Bz centroid.

The binding of Na^+, Cu^+, or Ag^+ with Bz brings about IR spectral changes of the Bz moiety. Figure 4 shows the MP2/aVDZ IR spectra of the $TM^{n+}-Bz$ complexes as compared to the pure benzene. As we discussed earlier, for the pure Bz, a CH out-of-plane bending mode appears at 678 cm^{-1}, a CH in-plane bending mode at 1052 cm^{-1}, a ring distortion mode at 1469 cm^{-1}, and a CH stretching mode at 3229 cm^{-1}. For the Na^+-Bz complex, two low frequencies at 726 and 886 cm^{-1} are out-of-plane bending modes, which are blue-shifted. The symmetric ring breathing mode appears at 996 cm^{-1}, the in-plane H bending mode at 1046 cm^{-1}, and the ring distorting mode at 1452 cm^{-1}. The IR spectra of Cu^+-Bz and Ag^+-Bz complexes are similar to those of the Na^+-Bz complex, while the significant difference is that the C−H stretching mode is IR inactive for the Na^+-Bz complex, which is due partly to the polarization of C−H bond reduced by the interaction between Na^+ and benzene.

C2. Binding of Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} to Benzene. Because the effective ionic radii for octa-coordinated Li^+, Na^+, and K^+ are 0.92, 1.16, and 1.51 Å,[36] respectively, the electrostatic interaction contribution in the Bz complexes with an alkali metal cation decreases as the distance between the metal cation and the Bz centroid increases. However, such a trend is not observed for the Bz complexes with Cu^+, Ag^+, and Au^+. The effective ionic radii for hexa-coordinated Cu^+, Ag^+, and Au^+ are 0.77, 1.15, and 1.37 Å. Because of the strong electron affinity of Au^+, the binding of Au^+ with Bz is clearly stronger than those of Cu^+ and Ag^+. Unlike those of Na^+, Cu^+, and Ag^+, the charge transfer from Au^+ to Bz is significant (q(Bz) = 0.21). The NBO analysis shows that

Table 2. BSSE-Corrected B3LYP, PBE, MP2, QCISD, and CCSD(T) Binding Energies (kcal/mol) of the TM^{n+}−Bz Complexes[a]

	B3LYP/aVDZ			PBE/aVDZ			MP2/aVDZ			QCISD/6-31G**		
	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$
$^1(Na-Bz)^+$	**20.4**			**24.3**			**21.7**			**23.8**		
$^1(Cu-Bz)^+$	**50.9**	**50.4**	**50.7**	63.6	63.6	63.6	**52**.2			**44.1**	**44.1**	**44.1**
$^1(Ag-Bz)^+$	34.5	**38.8**	**38.9**	40.9	**45.4**	**45.8**	**35.1**			31.0	31.8	**31.9**
$^1(Au-Bz)^+$	41.4	**58.9**	**58.4**	50.3	**69.7**	**69.7**	47.9	61.2	**61.7**	29.5	**46.5**	**46.2**
$^1(Pd-Bz)^{2+}$		**186.8**	**186.9**		213.3	**213.9**		152.5	**164.6**		159.5	**162.8**
$^1(Pt-Bz)^{2+}$		**203.3**	**203.7**		228.1	**231.1**		175.6	**181.8**		173.2	**177.1**
$^1(Hg-Bz)^{2+}$	128.1	**150.8**	148.4	143.1	**163.6**	161.5	114.1	**122.2**	**121.9**	108.4	**120.0**	117.8
$^3(Pd-Bz)^{2+}$	**198.5**			**218.4**			**164.7**			**169.1**		
$^3(Pt-Bz)^{2+}$	**200.0**			**224.9**			**177.9**			**172.2**		

	MP2/aVTZ//MP2/aVDZ			MP2/CBS//MP2/aVDZ			CCSD(T)/aVDZ//MP2/aVDZ			CCSD(T)/CBS//MP2/aVDZ		
	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$	$\pi_{cen}(\eta^6)$	$\pi_c(\eta^3)$	$\pi_{off}(\eta^4)$
$^1(Na-Bz)^+$	**21.7**			**21.7**			**22.1**			**22.1**		
$^1(Cu-Bz)^+$	**59.7**			**62.9**			**44.1**			**56.0**		
$^1(Ag-Bz)^+$	**40.9**			**43.5**			**31.4**			**40.1**		
$^1(Au-Bz)^+$	55.9	68.9	**69.7**	52.9	72.1	**73.1**	39.6	52.7	**52.9**	50.9	63.6	**64.3**
$^1(Pd-Bz)^{2+}$		190.5	**205.5**		206.5	**222.7**		150.1	**157.4**		204.1	**215.4**
$^1(Pt-Bz)^{2+}$		221.8	**230.7**		241.3	**251.3**		171.7	**173.8**		237.4	**243.4**
$^1(Hg-Bz)^{2+}$	134.4	**143.8**	142.9	142.9	**152.8**	151.7	113.4	**123.0**	121.9	142.3	**153.7**	151.7
$^3(Pd-Bz)^{2+}$	**198.4**			**212.6**			**165.4**			**213.3**		
$^3(Pt-Bz)^{2+}$	**219.1**			**236.4**			**174.3**			**232.8**		

	lowest structure MP2/aVDZ	MP2/aVTZ/ MP2/aVDZ	MP2/CBS//MP2/aVDZ	CCSD(T)/aVDZ/ MP2/aVDZ	CCSD(T)/CBS//MP2/aVDZ	expt.
$^1(Na-Bz)^+$	$\pi_{cen}(\eta^6)$	**21.7**	**21.7**	**22.1**	**22.1**	22.1 ± 1.4
$^1(Cu-Bz)^+$	$\pi_{cen}(\eta^6)$	**59.7**	**62.9**	**45.4**	**56.0**	52 ± 5
$^1(Ag-Bz)^+$	$\pi_{cen}(\eta^4)$	**40.9**	**43.5**	**31.4**	**40.1**	37.4 ± 1.7
$^1(Au-Bz)^+$	$\pi_{off}(\eta^4)$	**69.7**	**73.1**	**52.9**	**64.3**	69 ± 7
$^1(Pd-Bz)^{2+}$	$\pi_{off}(\eta^4)$	**205.5**	**222.7**	**157.4**	**215.4**	
$^1(Pt-Bz)^{2+}$	$\pi_{off}(\eta^4)$	**230.7**	**251.3**	**173.8**	**243.4**	
$^1(Hg-Bz)^{2+}$	$\pi_c(\eta^3)$	**143.8**	**152.8**	**123.0**	**153.7**	
$^3(Pd-Bz)^{2+}$	$\pi_{cen}(\eta^6)$	**198.4**	**212.6**	**165.4**	**213.3**	
$^3(Pt-Bz)^{2+}$	$\pi_{cen}(\eta^6)$	**219.1**	**236.4**	**174.3**	**232.8**	

[a] The energy was corrected by ZPE. For MP2/aVTZ//MP2/aVDZ, MP2/CBS//MP2/aVDZ, CCSD(T)/aVDZ/MP2/aVDZ, and CCSDT(T)/CBS//MP2/aVDZ calculations, only the binding energies of the most stable structures are collected in this table. For $^1(Cu-Bz)^+$ and $^1(Au-Bz)^+$, only the $\pi_{cen}(\eta^6)$ conformation is stable at the MP2/aVDZ level (ref 10d). Experimental values are from refs 30−32. For the TM^{2+}−Bz complexes, the spin−orbit coupling corrections for the binding energy by using the multiconfiguration calculations are no more than 0.01 kcal/mol due to the calcellation of errors, which is not included in this table.

Table 3. Calculated Vibrational Frequencies of Benzene[a]

approx. mode	B3LYP	MP2	QCISD	expt.
CH: out-of-plane bending	413_0	434_0	399_0	
CH: out-of-plane bending	679_{112}	746_{116}	685_{112}	673
ring stretching	974_0	1066_0	976_0	992
ring breathing	982_0	967_0	1012_0	
CH; in-plane bending	1024_6	1010_5	1057_4	1038
D-ring: ring distrortion	1448_6	1411_5	1491_5	1486
CC: stretching[b]	1584_0	1560_0	1647_0	
CH: stretching	3095_{35}	3101_{28}	3210_{36}	3210

[a] The aVDZ basis set was employed. The experimental values are from ref 33. Frequencies are in cm^{-1}, and IR intensities in km/mol are in subscripts. [b] IR inactive; not observed in the experiment (ref 34).

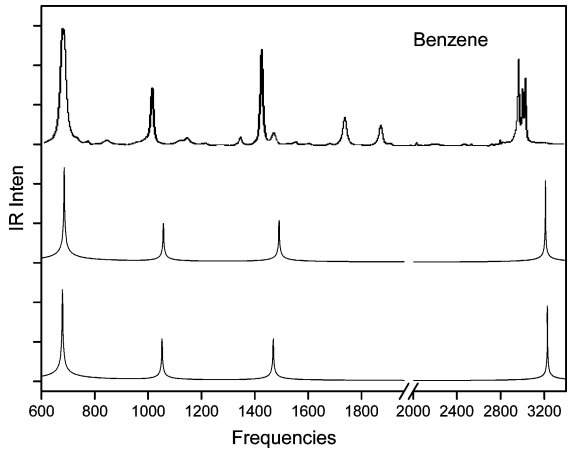


Figure 3. IR spectra of benzene [MP2/aVDZ (bottom), QCISD/6-31G** (middle), experiment (top)]. Experimental spectra are from ref 34.

the occupation of "6s" AO of Au^+ in the π_{cen} conformer is much smaller than the π_{off} and π_C conformers, indicating that the TM←π donation from Bz to the unoccupied "s" orbital of Au is less favorable in the π_{cen} structure.

For the singlet states of the Pd^{2+}−Bz and Pt^{2+}−Bz complexes, the B3LYP calculations predict that the π_{off} and π_C structures are isoenergetic, whereas the MP2 calculations show that the π_{off} conformer is particularly more stable. For the triplet states of Pd^{2+} and Pt^{2+}, there are two singly occupied AOs with no empty d orbitals. B3LYP, PBE, and CISD predict that in the case of Pd^{2+}−Bz complex the triplet π_{cen} structure is 12 kcal/mol more stable than the singlet π_{off} and π_C structures. On the other hand, MP2/CBS and CCSD(T)/CBS calculations show that for both Pd^{2+}−Bz and Pt^{2+}−Bz complexes the singlet π_{off} (η^4, which is practically similar in shape to η^2) structures are 17−19 and 24−31 kcal/mol more stable, respectively, than the triplet π_{cen} structures, which agrees with the observed singlet state of benzene η^2 coordinated with Pt^{2+} complexes.[15,16] Thus, MP2 calculations are likely to be more reliable than B3LYP, PBE, and CISD, while of course the CCSD(T)/CBS results would be the most reliable. Nevertheless, for more reliable results, the singlet−triplet separation would require multiconfiguration studies, which need to be done in the future. At the CCSD(T)/CBS level, the singlet−triplet splitting is 27.9 kcal/mol for Pd^{2+} and 25.4 kcal/mol for Pt^{2+} (or 30.6 kcal/mol for Pd^{2+} and 30.8 kcal/mol for Pt^{2+} after weighted averaging over all of the angular momentum states J:[10d] $E_{avg} = \sum_J[(2J + 1)/((2S + 1)(2L + 1))]E_J$). However, it may be energetically favorable that an electron promotes from a singly occupied d orbital to another in Pd^{2+} or Pt^2, which makes it possible a TM←ligand donation from an occupied π orbital or a lone electron pair to an empty d orbital of Pd^{2+} or Pt^{2+} (Figure 5a,b). After Pd^{2+} and Pt^{2+} interact with Bz, their single states are 2.1 and 10.6 kcal/mol more stable than their respective triplet states, respectively (Table 2).

Molecular dications formed by attachment of a TM^{2+} dication to Bz show significant charge transfer, as in Table 4. Although the binding of Pd^{2+}, Pt^{2+}, or Hg^{2+} with Bz is very strong, the intermolecular mode frequency is very small (Table 4), which indicates that the large binding energy of TM^{2+} attached to Bz is mainly attributed to the electrostatic interaction, but not covalent bonding. However, the TM←π donation makes the binding of Pd^{2+}/Pt^{2+} favor the π_{off} structure (Figure 5a,b). Because the s−d orbital energy splitting of Hg is very large, the s−d hybridization is not favorable for Hg^{2+}, and the "5d" AOs of Hg^{2+} are almost fully occupied. For the Hg^{2+}−Bz complex, the occupation of "6s" AO on the Hg atom indicates the TM←π donation from a π MO of Bz to an unoccupied "s" AO of Hg^{2+} (Figure 5c). As a result, the TM←π donation from Bz to the unoccupied "6s" AO of Hg^{2+} favors the π_C or π_{off} structure. The NBO analysis also shows that the occupation of "6s" AO for Hg^{2+} in the π_{cen} structure is much smaller than that of the π_C or π_{off} structure.

The π complexes of Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} give a few split IR peaks for the C−C stretching modes at about 1600 cm^{-1}, and several split IR peaks of the C−H stretching modes. The degenerate C−C stretching vibrations of the pure Bz are IR inactive at 1624 cm^{-1}, while they split into two IR active frequencies at 1576 and 1606 cm^{-1} for the Au^+−Bz complex. Because of the delocalization of the π ring, the C−C stretching frequencies are red-shifted as compared to those of Bz. The degenerate carbon ring distortion frequencies of Bz also split into two frequencies, 1446 and 1490 cm^{-1}, for the Au^+−Bz complex. The situation is similar for the π complexes of Pd^{2+}, Pt^{2+}, and Hg^{2+}. For Pd^{2+}−Bz, the C−C bond is greatly lengthened. The frequency of the related C−C stretching is also greatly red-shifted. The IR spectrum of the Pt^{2+}−Bz complex is similar to that of the Pt^{2+}−Bz complex. However, the Pt^{2+}−Bz complex gives the C−C stretching mode frequency (1363 cm^{-1}). Furthermore, more C−H stretching modes are IR active for Au^+−, Pd^{2+}−, Pt^{2+}−, and Hg^{2+}−Bz complexes, while the Hg^{2+}−Bz complex gives more red-shift in the C−H stretching mode frequency, which is probably correlated to the σ coordination. As shown in Figure 4, the symmetrical C−H stretching vibrational frequencies are split for the cases of the Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} complexes, and, in particular, the splitting for the Hg^{2+} complex is very wide.

IV. Conclusion

We have investigated the conformations and interaction energies of Bz with TM^{n+} (Na^+, Cu^+, Ag^+, Au^+, Pd^{2+},

Table 4. Intermolecular Mode Frequencies and Charge Transfer (CT) of the Lowest Energy Structures of the TM^{n+}−Bz Complexes[a]

	$^1(Na-Bz)^+$	$^1(Cu-Bz)^+$	$^1(Ag-Bz)^+$	$^1(Au-Bz)^+$	$^1(Pd-Bz)^{2+}$	$^1(Pt-Bz)^{2+}$	$^1(Hg-Bz)^{2+}$
ω_i	193_{35}	224_2	186_6	297_0	278_{11}	379_{36}	252_0
d	241	170	212	202	163	170	220
CT	0.0	−0.04	0.02	0.21	1.18	1.11	0.63

[a] Frequencies (ω_i in cm^{-1}), IR intensities (in km/mol in subscripts), and vertical distances from a cation to the Bz plane (d in pm) were obtained at the MP2/aVDZ level. The amount of CT was obtained in NBO charges at the MP2/aVTZ level.

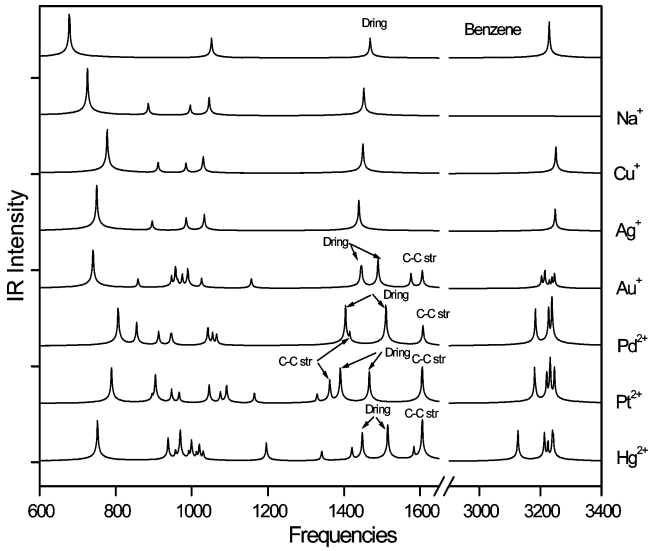


Figure 4. MP2/aVDZ calculated IR spectra of the pure Bz and the TM^{n+}−Bz complexes.

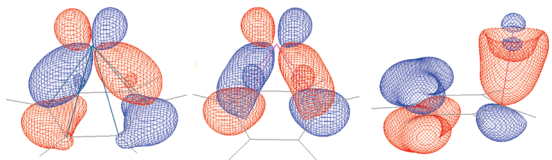


Figure 5. Occupied molecular orbitals of the TM^{2+} dication attached to Bz.

Pt^{2+}, and Hg^{2+}) using B3LYP/aVDZ, PBE/aVDZ, QCISD/6-31G**, MP2/aVDZ, and MP2/aVTZ calculations and MP2/CBS and CCSD(T)/CBS approximations. The CCSD(T)/CBS results agree with the experimental values, in particular, in the binding energies and singlet−triple energy separation. The QCISD/6-31G** binding energies of Na^+−Bz, Cu^+−Bz, and Ag^+−Bz also agree well with the experimental values, while the QCISD calculations apparently underestimate that of Au^+−Bz. For TM^{n+}−Bz (TM^{n+} = Pd^{2+}, Pt^{2+}), MP2 calculations show that the singlet π conformers are more stable than the triplet π conformers. Although the binding of Cu^+/Ag^+ with Bz arises mainly from the electrostatic and induction interactions, the binding of Au^+, Pd^{2+}, Pt^{2+}, or Hg^{2+} with Bz involves significant charge transfer, and the electrostatic interaction still plays a more important role than the covalent bonding. Because of the characteristic binding, the binding sites of the metal cation in the Au^+−, Pd^{2+}−, Pt^{2+}−, and Hg^{2+}−Bz complexes depend on the metal. The Na^+−, Cu^+−, and Ag^+−Bz complexes prefer the $\pi_{cen}(\eta^6)$ structure of C_{6v} symmetry, The Na^+ complex prefers the $\pi_{cen}(\eta^6)$ structure of C_{6v} symmetry. The Cu^+ and Ag^+ cations can be over all of the benzene plane, but would still favor the $\pi_{cen}(\eta^6)$ structure of C_{6v} symmetry. The Au^+−Bz complex favors the $\pi_{off}(\eta^2)$ or $\pi_C(\eta^1)$ structure, and the Pd^{2+}− and Pt^{2+}−Bz complexes favor the $\pi_{off}(\eta^4)$ structures. The Hg^{2+}−Bz complex favors the σ or $\pi_C(\eta^1)$ structure.

The TM^{n+}−Bz complexes give characteristic IR peaks. The MP2/aVDZ calculations show that for the Na^+, Cu^+, and Ag^+ complexes, the out-of-plane bending frequencies are blue-shifted with respect to that of the pure Bz at 679 cm^{-1}, and an out-of-plane bending mode and a symmetric ring breathing mode appear at 800−1000 cm^{-1}. In contrast to the nonactive C−C stretching mode of the pure Bz and the Na^+, Cu^+, and Ag^+ complexes due to the C_6 symmetry, the C−C stretching modes in the Au^+, Pd^{2+}, Pt^{2+}, and Hg^{2+} complexes are IR active due to the nonsymmetric geometry. This distinction would be useful to identify the complex symmetry and find the binding site of the metal on the benzene. The Au^+−Bz complex shows two IR active frequencies at 1576 and 1606 cm^{-1}. For the Pd^{2+}− and Pt^{2+}−Bz complexes, the C−C bond is lengthened, resulting in a large red-shift. The Hg^{2+}−Bz complex gives a large red-shift in the C−H stretching frequency due to the σ coordination.

Acknowledgment. K.S.K. acknowledges the support from KOSEF (WCU, R32-2008-000-10180-0; EPB Center, R11-2008-052-01000), BK21 (KRF), and GRL (KICOS). H.-B.Y. acknowledges the support by the 985 Project of Hunan University (China). H.M.L. acknowledges the KRF Grant of MOEHRD (Korea) (KRF-2006-353-C00022). The support from the KISTI Supercomputing Center (KSC-2008-K08-0002, KSC-2007-S00-3005) is acknowledged.

References

(1) (a) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303–1324. (b) Kumpf, R.; Dougherty, D. A. *Science* **1993**, *261*, 1708–1710. (c) Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A. *J. Am. Chem. Soc.* **1996**, *118*, 2307–2308. (d) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145–4186. (e) Pullman, A.; Berthier, G.; Savinelli, R. *J. Comput. Chem.* **1997**, *18*, 2012–2022. (f) Kim, K. S.; Lee, J. Y.; Lee, S. J.; Ha, T.-K.; Kim, D. H. *J. Am. Chem. Soc.* **1994**, *116*, 7399–7400. (g) Pierpont, C. G.; Buchanan, R. M. *Coord. Chem. Rev.* **1981**, *38*, 45–87. (h) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.

(2) (a) Cabarcos, O. M.; Weinheimer, C. J.; Lisy, J. M. *J. Chem. Phys.* **1999**, *110*, 8429–8435. (b) Cabarcos, O. M.; Weinheimer, C. J.; Lisy, J. M. *J. Chem. Phys.* **1998**, *108*, 5151–5154.

(3) Rodgers, M. T.; Armentrout, P. B. *Mass Spectrom. Rev.* **2000**, *19*, 215–247.

(4) Kim, D.; Hu, S.; Tarakeshwar, P.; Kim, K. S.; Lisy, J. M. *J. Phys. Chem. A* **2003**, *107*, 1228–1238.

(5) (a) Schröder, D.; Wesendrup, R.; Hertwig, R. H.; Dargel, T. K.; Grauel, H.; Koch, W.; Bender, B. R.; Schwarz, H. *Organometallics* **2000**, *19*, 2608–2615. (b) Kim, D.; Lee, E. C.; Kim, K. S.; Tarakeshwar, P. *J. Phys. Chem. A* **2007**, *111*, 7980–7986. (c) Lee, J. Y.; Lee, S. J.; Choi, H. S.; Cho, S. J.; Kim, K. S.; Ha, T. K. *Chem. Phys. Lett.* **1995**, *232*, 67–71.

(6) (a) Hu, J.; Barbour, L. J.; Gokel, G. W. *J. Am. Chem. Soc.* **2001**, *123*, 9486–9487. (b) Gokel, G. W.; Barbour, L. J.; Ferdani, R.; Hu, J. *Acc. Chem. Res.* **2002**, *35*, 878–886. (c) Choi, H. S.; Suh, S. B.; Cho, S. J.; Kim, K. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12904–12099. (d) McFail-Isom, L.; Shui, X. Q.; Williams, L. D. *Biochemistry* **1998**, *37*, 17105–17111. (e) Zaric, S. D.; Popovic, D. M.; Knapp, E. W. *Chem.-Eur. J.* **2000**, *6*, 3935–3942.

(7) (a) Hong, B. H.; Lee, J. Y.; Lee, C.-W.; Kim, J. C.; Bae, S. C.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, *123*, 10748–10749. (b) Hong, B. H.; Bae, S. C.; Lee, C.-W.; Jeong, S.; Kim, K. S. *Science* **2001**, *294*, 348–351. (c) Singh, N. J.; Lee, E. C.; Choi, Y. C.; Lee, H. M.; Kim, K. S. *Bull. Chem. Soc. Jpn.* **2007**, *80*, 1437–1450. (d) Singh, N. J.; Lee, H. M.; Suh, S. B.; Kim, K. S. *Pure Appl. Chem.* **2007**, *79*, 1057–1075.

(8) (a) Oh, K. S.; Lee, C.-W.; Choi, H. S.; Lee, S. J.; Kim, K. S. *Org. Lett.* **2000**, *2*, 2679–2681. (b) Choi, H. S.; Kim, D.; Tarakeshwar, P.; Suh, S. B.; Kim, K. S. *J. Org. Chem.* **2002**, *67*, 1848–1851. (c) Yun, S.; Kim, Y.-O.; Kim, D.; Kim, H. G.; lhm, H.; Kim, J. K.; Lee, C.-W.; Lee, W. J.; Yoon, J.; Oh, K. S.; Yoon, J.; Park, S.-M.; Kim, K. S. *Org. Lett.* **2003**, *5*, 471–474. (d) Singh, N. J.; Lee, H. M.; Hwang, I.-C.; Kim, K. S. *Supramol. Chem.* **2007**, *19*, 321–332.

(9) (a) Gapeev, A.; Dunbar, R. C. *J. Am. Chem. Soc.* **2001**, *123*, 8360–8365. (b) Dunbar, R. C. *J. Phys. Chem. A* **2000**, *104*, 8067–8074. (c) Dunbar, R. C. *J. Phys. Chem. A* **1998**, *102*, 8946–8952. (d) Tsuzuki, S.; Yoshida, M.; Uchimaru, T.; Mikami, M. *J. Phys. Chem. A* **2001**, *105*, 769–773. (e) Alberti, M.; Aguilar, A.; Lucas, J. M.; Lagana, A.; Pirani, F. *J. Phys. Chem. A* **2007**, *111*, 1780–1787. (f) Ruan, C. H.; Rodgers, M. T. *J. Am. Chem. Soc.* **2004**, *126*, 14600–14610.

(10) (a) Polfer, N. C.; Oomens, J.; Morre, D. T.; von Helden, G.; Meijer, G.; Dunbar, R. C. *J. Am. Chem. Soc.* **2006**, *128*, 517–525. (b) Dunbar, R. C.; Moore, D. T.; Oomens, J. *J. Phys. Chem. A* **2006**, *110*, 8316–8326. (c) Moore, D. T.; Oomens, J.; Eyler, J. R.; von Helden, G.; Meijer, G.; Dunbar, R. C. *J. Am. Chem. Soc.* **2005**, *127*, 7243–7254. (d) Yi, H.-B.; Diefenbach, M.; Choi, Y. C.; Lee, E. C.; Lee, H. M.; Hong, B. H.; Kim, K. S. *Chem.-Eur. J.* **2006**, *12*, 4885–4892. (e) Pandey, R.; Rao, B. K.; Jena, P.; Blanco, M. A. *J. Am. Chem. Soc.* **2001**, *123*, 3799–3808. (f) Roszak, S.; Balasubramanian, K. *Chem. Phys. Lett.* **1995**, *234*, 101–106.

(11) Nazin, G. V.; Qiu, X. H.; Ho, W. *Science* **2003**, *302*, 77–81.

(12) Dargel, T. K.; Hertwig, R. H.; Koch, W. *Mol. Phys.* **1999**, *96*, 583–592.

(13) Petrie, S.; Radom, L. *Int. J. Mass Spectrom.* **1999**, *192*, 173–183.

(14) Alcamí, M.; González, A. I.; Mó, O.; Yáñez, M. *Chem. Phys. Lett.* **1999**, *307*, 244–252.

(15) Johansson, L.; Tilset, M.; Labinger, J. A.; Bercaw, J. E. *J. Am. Chem. Soc.* **2000**, *122*, 10486–10487.

(16) (a) Norris, C. M.; Reinartz, S.; White, P. S.; Templeton, J. L. *Organometallics* **2002**, *21*, 5649–5656. (b) Reinartz, S.; White, P. S.; Brookhart, M.; Templeton, J. L. *J. Am. Chem. Soc.* **2001**, *123*, 12724–12725.

(17) Siu, F. M.; Ma, N. L.; Tsang, C. W. *J. Am. Chem. Soc.* **2001**, *123*, 3397–3398.

(18) (a) Nechaev, M.; Rayón, V.; Frenking, G. *J. Phys. Chem. A* **2004**, *108*, 3134–3142. (b) Lupinetti, A. J.; Fau, S.; Frenking, G.; Strauss, S. H. *J. Phys. Chem. A* **1997**, *101*, 9551–9559.

(19) (a) Sievers, M. R.; Jarvis, L. M.; Armentrout, P. B. *J. Am. Chem. Soc.* **1998**, *120*, 1891–1899. (b) Corral, I.; Mo, O.; Yanez, M. *J. Phys. Chem. A* **2003**, *107*, 1370–1376.

(20) Pople, J. A.; Head-Gordon, M. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(21) (a) Bergner, A.; Dolg, M.; Kuechle, W.; Stoll, H.; Preuss, H. *Mol. Phys.* **1993**, *80*, 1431–1441. (b) Kaupp, M.; Schleyer, P. v. R.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1991**, *94*, 1360–1366. (c) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866–872. (d) Dolg, M.; Stoll, H.; Preuss, H.; Pitzer, R. M. *J. Phys. Chem.* **1993**, *97*, 5852–5859.

(22) Feller, D.; Glendening, E. D.; de Jong, W. A. *J. Chem. Phys.* **1999**, *110*, 1475–1491.

(23) Matin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408–3420.

(24) (a) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646. (b) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208–1221.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03*, revision B.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

(26) Lee, S. J.; Chung, H. Y.; Kim, K. S. *Bull. Korean Chem. Soc.* **2004**, *25*, 1061–1064.

(27) (a) Lide, D. R., Ed. *Handbook of Chemistry and Physics*, 1992; 73rd ed., CRC Press Ltd.: Boca Raton, FL, pp 10−211. (b) Bilodeau, R. C.; Scheer, M.; Haugen, H. K. *J. Phys. B: At. Mol. Opt.* **1998**, *31*, 3885–3891. (c) Gantefor, G.; Kraus, S.; Eberhardt, W. *J. Electron Spectrosc. Relat. Phenom.* **1998**, *88*, 35–40. (d) Scheer, M.; Brodie, C. A.; Bilodeau, R. C.; Haugen, H. K. *Phys. Rev. A* **1998**, *58*, 2051–2062. (e) Bilodeau, R. C.; Scheer, M.; Haugen, H. K.; Brooks, R. L. *Phys. Rev. A* **2000**, *61*, 012505-1–012505-7.

(28) Solcà, N.; Dopfer, O. *Angew. Chem., Int. Ed.* **2002**, *41*, 3628–3631.

(29) (a) Nicholas, J. B.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **1999**, *103*, 1394–1400. (b) Feller, D. *Chem. Phys. Lett.* **2000**, *322*, 543–548. (c) Pullman, A.; Berthier, G.; Savinelli, R. *J. Mol. Struct. (THEOCHEM)* **2001**, *537*, 163–172.

(30) Amicangelo, J. C.; Armentrout, P. B. *J. Phys. Chem. A* **2000**, *104*, 11420–11432.

(31) (a) Meyer, F.; Khan, F. A.; Armentrout, P. B. *J. Chem. Soc.* **1995**, *117*, 9740–9748. (b) Chen, Y.-M.; Armentrout, P. B. *Chem. Phys. Lett.* **1993**, *210*, 123–128. (c) Rodgers, M. T.; Armentrout, P. B. *J. Phys. Chem. A* **1997**, *101*, 1238–1249. (d) Andersen, A.; Muntean, F.; Walter, D.; Rue, C.; Armentrout, P. B. *J. Phys. Chem.* **2000**, *104*, 692–705.

(32) Guo, B. C.; Purnell, J. W.; Castleman, A. W., Jr. *Chem. Phys. Lett.* **1990**, *168*, 155–160.

(33) (a) Jaeger, T. D.; van Heijnsbergen, D.; Klippenstein, S. J.; von Helden, G.; Meijer, G.; Duncan, M. A. *J. Am. Chem. Soc.* **2004**, *126*, 10981–10991. (b) van Heijnsbergen, D.; von Helden, G.; Meijer, G.; Maitre, P.; Duncan, M. A. *J. Am. Chem. Soc.* **2002**, *124*, 1562–1563.

(34) Kinugasa, S.; Tanabe, K.; Tamura, T. *Spectral Database for Organic Compounds*; National Institute of Advanced Industrial Science and Technology (AIST): Japan.

(35) Turner, D. W.; Baker, C.; Baker, A. D.; Brundle, C. R. *Molecular Photoelectron Spectroscopy*; Wiley: New York, 1970.

(36) Shannon, R. D. *Acta Crystallogr.* **1976**, *A32*, 751–767.

CT900154X