JCTC Journal of Chemical Theory and Computation

Electrostatically Embedded Many-Body Expansion for Simulations

Erin E. Dahlke and Donald G. Truhlar*

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

Received August 29, 2007

Abstract: We have applied the electrostatically embedded many-body (EE-MB) method truncated at the two-body level (also called the pairwise additive EE-MB method or the EE-PA approximation) and the three-body level (called EE-3B) to calculate the gradient of the potential energy for a simulation box containing 64 water molecules. We employed the B3LYP density functional with the 6-31+G(d,p) basis set for this test case. We found that the EE-PA method is able to reproduce the magnitude of the gradient from a B3LYP/6-31+G(d,p) calculation on the entire system to within 1.0% with a 1.3% error for the maximum component of the gradient. Furthermore, the EE-3B method is able to reproduce the magnitude of the gradient to within 0.1% with a 0.2% error for the maximum component of the gradient. The good performance of the EE-MB methods for calculating forces and the highly parallel nature of these methods make them well suited for use in molecular dynamics simulations. Furthermore, since the methods can be used for efficient and accurate calculations of forces with any level of electronic structure theory that has analytic gradients and with any electronic structure package that allows for the presence of a field of point charges, these methods can readily be used with a wide variety of density functional theory and wave function theory methods.

Molecular simulations that use molecular mechanics potentials or other analytic potentials for the potential energy surface and classical mechanics for the nuclear motion have been carried out for nearly 50 years,[1] but although molecular mechanics potentials may give good agreement with experiment for the physical properties against which they are parametrized, they often give poor results when applied to properties outside this set. As a result of this shortcoming, molecular mechanics potentials must be developed anew or revalidated for each new system of interest and even for each property one wants to study. In the interest of developing more robust methods for calculating potential energies for molecular simulations, there is great interest in the direct use of quantum mechanical methods without analytic representations, i.e., direct dynamics. In particular, a quantum mechanical theoretical model chemistry[2−4] can be validated against a broad data set for predicting potential energy surfaces or properties dependent on them; and if the validation test is sufficiently broad, the quantum mechanical model chemistry is likely to have better predictive value than molecular mechanics because it more fully incorporates the relevant physics.

Due to the large system sizes for most condensed-phase simulations, even when using periodic boundary conditions,[5,6] model chemistries based on wave function theory[3,4] (WFT) such as second-order Møller−Plesset perturbation theory (MP2),[7] coupled cluster theory with single and double excitations (CCSD),[8] or CCSD with quasiperturbative triples[9] (CCSD(T)) are currently impractical in their original formulations, in part because of the rapid scaling in cost of these methods with respect to system size. (MP2, CCSD, and CCSD(T) scale as N^5, N^6, and N^7, respectively, where N is the number of atoms in the system.[10]) As a result, most direct dynamics simulations are carried out using density functional theory[11] (DFT), whose scaling cost, with popular algorithms, increases only as N^3 or N^4. Due to the quantum mechanical nature of DFT, these simulations are significantly more expensive[12,13] than their counterparts with molecular mechanics or analytic potentials, but the added cost is rationalized in the hope that the energies obtained are much more accurate and the functionals are more transferable. A drawback to such conventional calculations is that only a relatively small number of density functionals have been implemented in the most efficient periodic-boundary-condition simulation packages, and when a newer, more accurate kind of functional becomes available, it may require specialized programming to be made available in efficient packages.

In recent years several groups have emphasized the advantage of many-body expansions[14−29] and other fragmentation methods[30−39] for calculations on large systems. A crucial aspect of using any such method for geometry optimization or for calculating forces or molecular dynamics is the ability to formulate efficient algorithms for analytic gradients of the potential energy surface. The pioneering fragment molecular orbital (FMO) method[14−18,22−24,27,28] has been particularly successful for large systems, especially proteins, and methods were developed for nearly analytic

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

10.1021/ct700223r CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/12/2007

restricted Hartree−Fock (RHF) gradients[15] and analytic derivatives of the two-body electrostatic interactions between widely separated fragments.[27] The molecular fractionation with conjugate caps method,[31−34] which does not include three-body or higher-order terms or long-range electrostatics, but rather simulates the local chemical environment of fragments with conjugate caps, has also been applied very successfully to proteins and allows[33] for convenient calculation of dimer gradients. The method has been extended to include long-range electrostatic fields both with[35] and without[36] truncation, and in the former case gradients were obtained.

We have formulated an efficient and accurate many-body expansion method in a way that yields computationally efficient energy gradients for all electronic structure levels for which they are available for the fragments,[25] and in this letter we test the accuracy of the gradients and describe the applicability to molecular simulations of this new approach, which is called the electrostatically embedded many-body (EE-MB) expansion. The EE-MB method can be used with both wave function methods such as MP2 and CCSD(T) and with DFT. For both types of methods it makes the scaling more manageable, and it has the distinct advantage that it can be used in conjunction with any electronic structure package (allowing researchers to utilize any WFT level or any density functional of their choosing). Very accurate results can be obtained in the three-body approximation[25,26,29] with a scaling of N^3. The EE-MB method is very general and can be applied to molecular liquids such as simulations of aqueous solutions or (when extended to include a scheme, such as link atoms[40−43] or conjugated caps,[32] for terminating fragments at fragment boundaries that pass through bonds) to large covalent systems such as polymers or proteins. In this work we focus on its utility for simulating molecular liquids and use pure water as an example. A key issue is that analytic gradients are available for three-body and higher-order terms as well as two-body terms and for near as well as far fragments, while retaining the key advantage that the electrostatic field of the rest of the system is not truncated.

The complete details of the EE-MB method are presented elsewhere,[25] and so here we present only a brief overview of the method. For any level of theory (e.g., MP2 or CCSD(T) with a given basis set, or DFT with a given functional and basis) we can expand the potential energy of a system of N monomers (where a monomer can be a single molecule a small collection of molecules) in a many-body expansion given by

$$V = V_1 + V_2 + V_3 + \cdots + V_N \tag{1}$$

where V_n is the n-body term. Truncating at V_2 is called the pairwise additive approximation (PA), and truncating at V_3 is called the three-body (3B) approximation. For a system with N monomers, V_1 involves calculating all N monomer energies, V_2 involves calculating $(N(N-1))/2$ dimer energies, and V_3 involves calculating $(N(N-1)(N-2))/3!$ trimer energies. If the n-mer calculations are performed in vacuum one has a conventional many-body expansion; however, in the EE-MB methods (where MB = PA or 3B) the n-mer calculations are performed in a field of point charges at the nuclear positions of the $N - n$ missing monomers.

The applicability of many-body expansion methods to Monte Carlo simulations has been discussed by Christie and Jordan,[21] and so in this work we will focus on application to molecular dynamics calculations. In previous work we have demonstrated the ability of the EE-MB methods to accurately reproduce the energetics of a series of water clusters ranging in size from 5 to 20 molecules.[25,26,29] In that work we found that the EE-PA method was able to reproduce the energy of a system to within 0.8% and that the EE-3B method was able to reproduce the energy to within 0.3%, and we also discussed the efficiency with which gradients could be calculated using the EE-MB method. Because the largest calculation carried out for these methods is a dimer (in the EE-PA method) or a trimer (in the EE-3B method) the problem of needing to carry out one very large calculation is reduced to carrying out a very large number of small calculations, which is more practical on most computers. In this way one also avoids the very high scaling of many WFT methods, such as CCSD(T), and this makes it possible to apply the EE-MB levels of theory to simulations of very large systems.

Within the EE-MB approximation the EE-PA and EE-3B energies can be written as

$$E_{\mathrm{EE-PA}} = \sum_{i>j} E_{ij} - (N-2)\sum_{i} E_i \tag{2}$$

$$E_{\mathrm{EE-3B}} = \sum_{i>j>k} E_{ijk} - (N-3)\sum_{i>j} E_{ij} + \frac{(N-3)(N-2)}{2}\sum_{i} E_i \tag{3}$$

where N is the number of particles in the system and where E_i, E_{ij}, and E_{ijk} are the energies of the embedded monomers, dimers, and trimers. Since the gradient is a linear operator it follows that

$$\nabla E_{\mathrm{EE-PA}} = \sum_{i<j}^{N} \nabla E_{ij} - (N-2)\sum_{i}^{N} \nabla E_i \tag{4}$$

and similarly

$$\nabla E_{\mathrm{EE-3B}} = \sum_{i>j>k} \nabla E_{ijk} - (N-3)\sum_{i>j} \nabla E_{ij} + \frac{(N-3)(N-2)}{2}\sum_{i} \nabla E_i \tag{5}$$

where analytic gradients are therefore available for any method that has analytic gradients for the monomer, dimer, and, in the case of the EE-3B method, trimer calculations, provided that the program allows for fractionally charged point charges as pseudonuclei. Since the magnitudes of the point charges are fixed in our EE-MB calculations, the point charges act like fractionally charged nuclei with no basis functions; therefore, as the system evolves during the course of a simulation there is no need to update the charges. Nevertheless, one should also note that all terms on the right-hand sides of eqs 4 and 5 contribute to all components of the gradient. For example, even if $m \neq i$, $m \neq j$, and $m \neq k$, one still has that ∇E_{ijk}, ∇E_{ij}, and ∇E_i all contribute to the gradient components corresponding to the coordinates of monomer m.

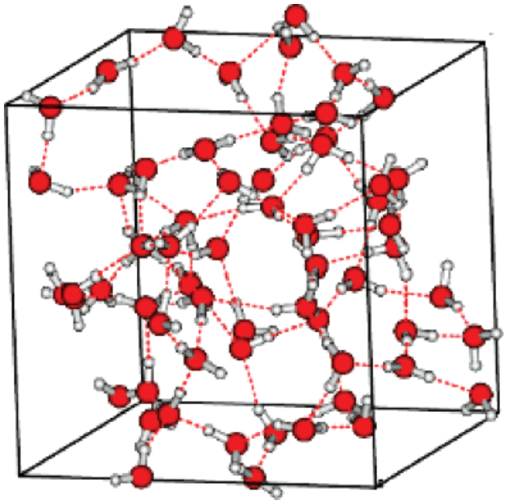


Figure 1. Simulation box used for single point gradient calculation.

As mentioned previously, while the application of the EE-MB method in this work is limited to a water cluster in which there are no covalent bonds present between the monomers, the form of eq 2 is very similar to the equation used to calculate protein−ligand interactions with the molecular fractionation with conjugate caps method presented in ref 33.

In order to demonstrate the ability of the EE-MB method with fixed point charges to yield accurate gradients, we have calculated a single-point gradient on a simulation box containing 64 water molecules (see Figure 1), without periodic boundary conditions, using both the EE-PA and EE-3B methods with the B3LYP[44−47] density functional and the 6-31+G(d,p)[48] basis set, and we have compared the results to a conventional B3LYP/6-31+G(d,p) calculation on the full system. (Although the EE-MB method can be used in conjunction with WFT methods such as MP2 and CCSD(T), we have limited ourselves for this validation test to the use of DFT because the rapid scaling of MP2 and CCSD(T) makes the calculation of a single-point gradient calculation on the full system, as required to test the EE-MB gradients, very expensive.) The work of Lenosky et al.[49] has shown that the use of a single gradient on a large system is a powerful tool for the optimization of methods, and therefore we use it here as a way to analyze the EE-MB method. Note that the gradient of a cluster of 64 water molecules provides 192 gradient components against which to test the EE-MB method.

The full calculation was carried out using the *Gaussian 03* software package.[50] The EE-PA and EE-3B calculations were carried out using the MBPAC 2007-2 software package.[51]

Table 1 compares the results of the EE-MB calculations to the gradient from the full B3LYP/6-31+G(d,p) calculation. Table 1 lists the magnitude of the gradient and the maximum component of the gradient from the EE-PA, EE-3B, and conventional B3LYP/6-31+G(d,p) calculations, the error in the gradient, the error in the maximum component of the gradient, and the mean absolute error in the components of the gradient predicted by the EE-PA and EE-3B methods. Table 1 also lists the percentage error in the magnitude of the gradient and in the maximum component of the gradient from the EE-PA, EE-3B calculations. From Table 1 it is clear that both the EE-PA and EE-3B methods are able to reproduce the forces for this system very well, with errors in the gradient of less than 0.0002 au (one atomic unit (au) of force equals one hartree per bohr) for the EE-PA method and less than 3×10^{-6} au for the EE-3B level, which corresponds to a percentage error of less than 1% for EE-PA and less than 0.01% for EE-3B. We see similarly good performance for the maximum component of the gradient, with the EE-PA method having an error of 1.3% and the EE-3B method having an error of 0.2%. The near-order-of-magnitude improvement as one goes from the EE-PA method to the EE-3B method is consistent with past studies[25,26,29] considering only energetics. The mean absolute error also shows that the EE-3B method performs better than EE-PA as it has a MAE of 4.07×10^{-4} au compared to a value of 6.23×10^{-4} au for the EE-PA method. The average magnitude of a component of the gradient is 1.37×10^{-2} au, so the mean absolute deviation as a percentage of the mean component is 4.5% for the EE-PA method and 3.0% EE-3B method.

These comparisons show that, even at the EE-PA level, the EE-MB method is able to achieve gradients in reasonable agreement with gradients calculated by conventional methods. One should be careful not to interpret the deviation as an error, just as the difference between conventional MP2 and conventional CCSD(T) is not an error but rather a difference between two model chemistries. In the present case the difference between conventional B3LYP and EE-3B/B3LYP is expected to be smaller than the difference of either from complete configuration interaction. A key issue is that eqs 4 and 5 provide an accurate theoretical model chemistry[2−4] with precise and convenient gradients. The deviation of EE-MB/DFT from conventional DFT will be of minor importance for many purposes, but the high precision of the gradients in the present algorithm will be a critical component of stable (nondrifting) molecular dynamics simulations.

Calculating the bulk properties of molecular liquids by the EE-MB method can be accomplished by employing periodic boundary conditions,[5,6] and this can be accomplished for EE-MB simulations by methods already developed for QM/MM simulations[52] augmented by a criterion to select the appropriate image of each monomer in the dimers and trimers. The latter can be accomplished by the nearest-image convention,[6] which is currently employed in simulations utilizing analytic functions for the potential energy functions. The nearest-image convention is widely used for pairwise potentials and has been modified[53] for three-body potentials, and its implementation is straightforward. Additionally, for any potential that decays more rapidly than R^{-3} (such as dispersion terms arising purely from quantum mechanical correlation) the use of a cutoff can be employed (typical cutoffs are one-half the box length for a cubic simulation box). In cases where one has long-range interactions, Ewald summations are used with molecular mechanics potentials to account for the interactions of point charges and dipolar molecules,[5] and they can be employed in the same way for the present electrostatic embedding terms.

The treatment of periodic images of the embedded quantum mechanical monomers, dimers, and trimer can be identical to methods employing periodic boundary in the context of combined quantum mechanical (QM) and mo-

Table 1. Errors in the Gradient and the Components of the Gradient (in Atomic Units) for the EE-PA and EE-3B Methods at the B3LYP/6-31+G(d,p) Level of Theory

	full	EE-PA[a]	EE-3B
magnitude of the gradient	1.8512×10^{-2}	1.8695×10^{-2}	1.8509×10^{-2}
max. component of the gradient	6.0892×10^{-2}	6.1686×10^{-2}	6.1005×10^{-2}
error			
magnitude of the gradient		1.8354×10^{-4}	-2.6305×10^{-6}
max. component of the gradient		7.9324×10^{-4}	1.1312×10^{-4}
% error			
magnitude of the gradient		0.99	−0.01
max. component of the gradient		1.30	0.19
MAE[b]		6.2257×10^{-4}	4.0667×10^{-4}

[a] The EE-MB calculations used point charges of −0.778 and 0.389 for oxygen and hydrogen atoms, respectively, as in past work.[25] [b] MAE denotes the mean absolute error (in atomic units) in the components of the gradient.

lecular mechanical (MM) calculations (QM/MM calculations[52,54−58]). The total energy for a QM/MM calculation can be written as

$$E(\mathrm{QM/MM}) = E(\mathrm{MM}) + E(\mathrm{QM}) + E(\mathrm{QM\text{-}MM}) \quad (6)$$

where E(MM) is the energy of the molecular mechanics system, E(QM) is the contribution from the quantum mechanical system, and E(QM-MM) is the contribution due to coupling of the MM and QM regions. For QM/MM methods that employ electronic embedding,[43,59−67] one of the terms in E(QM-MM) is computed along with E(QM) by calculating E(QM) in a field of molecular mechanics point charges. Therefore, each embedded monomer, dimer, or trimer calculation in an EE-MB calculation can be thought of as a simplified QM/MM calculation, in which the E(MM) term is zero (the interaction energy of the point charges is not included in the total EE-MB energy), and the only contribution to the E(QM-MM) term is from embedding the n-mer in an environment of point charges. There are a number[52,54,57,58,66,67] of examples in the literature in which periodic boundary conditions have been applied successfully to QM/MM calculations, and QM/MM codes can be used in conjunction with the EE-MB method by writing a subroutine to interface the existing code with the electronic structure package of one's choosing to carry out the EE-MB calculation. Furthermore, because all of the monomer, dimer, and trimer calculations are independent of each other, the EE-MB method is highly parallel, which allows for rapid energy calculations, even on very large systems.

Due to the expense of an accurate treatment of the electronic wave function near the nucleus of an atom, a variety of specialized approximations and procedures have been developed for plane-wave simulations of condensed-phase systems.[68−75] For example, ultrasoft pseudopotentials[68,70] are often used to keep the plane wave cutoff low, and such pseudopotentials must be carefully optimized to minimize inaccuracies.[71−73] Due to the small system sizes calculated in the EE-MB methods, all calculations can be carried out without pseudopotentials or with norm-conserving[76] (also called shape-consistent[77]) effective core potentials, which are less economical but more accurate. Also, well validated techniques developed for small-molecule calculations can be used.

It is interesting to consider the amount of time needed to carry out these kinds of calculations. Table 2 presents a series of hypothetical timings, for a calculation on 64 molecules, for methods that scale as aN^3 (e.g., BLYP, PBE, and M06-L), bN^4 (e.g., B3LYP, M06-2X), and cN^7 (e.g., MP4, CCSD(T)), where N is the number of atoms in the system, and a, b, and c are unknown prefactors, specific to each level of electronic structure theory. (It is an approximation to assume that this scaling holds for all N, including small N, but timing analyses are inherently approximate, and the present timing discussion is intended to illustrate scaling issues—not to be quantitative.) Table 2 shows that even on a single processor, the many-body approaches are far more cost-effective than conventional calculations. In this example, use of the EE-3B method would reduce the cost of a method that scales as cN^7 by 6 orders of magnitude on a system of 64 molecules, and use of the EE-PA method would reduce the cost by 7 orders of magnitude. Even for density functional theory it is clear from Table 2 that the EE-MB methods are cost-effective. For hybrid methods, such as B3LYP or M06-2X that scale as bN^4, both the EE-PA and EE-3B methods are less expensive than a conventional calculation on the full clusters. For nonhybrid methods the EE-PA method is an order of magnitude less expensive, and the EE-3B calculation is only a factor of 1.3 more expensive.

Table 2. Comparison of Hypothetical Timings for Full Calculations and EE-MB Calculations for a System Containing 64 Molecules

scaling	conventional	EE-PA	EE-3B
aN^3	$2.6 \times 10^5\ a$	$1.6 \times 10^4\ a$	$3.5 \times 10^5\ a$
bN^4	$1.7 \times 10^7\ b$	$3.2 \times 10^4\ b$	$7.0 \times 10^5\ b$
cN^7	$4.4 \times 10^{12}\ c$	$2.5 \times 10^5\ c$	$5.6 \times 10^6\ c$

All calculations in Table 2, both conventional and EE-MB, can be further speeded up by linear scaling algorithms,[18,52,78,79] but quantitative speedups depend strongly on the program and will not be estimated here. Nevertheless it is worthwhile to note that linear scaling can be achieved in EE-MB by using a cutoff to reduce the number of two-body or three-body terms that must be calculated. We showed that if a cutoff of 6 Å is used, then even for a cluster as small as $(H_2O)_{20}$ one can eliminate up to 44% of the pairs.[26] A key issue here is that the introduction of linear scaling is much simpler in the EE-MB approximation than in other methods of comparable accuracy because it simply involves limiting the number of dimers and/or trimers considered, but it does not require cutting off long-range electrostatics when they are treated by Ewald.

In summary, we have found that both the EE-PA and EE-3B methods are able to reproduce the gradient and the maximum component of the gradient for a simulation box containing 64 water molecules to within 1% and 0.1% respectively, at the B3LYP/6-31+G(d,p) level of theory. Probably more important though is the high-precision attainable when EE-MB methods are used as a theoretical model chemistry. Additionally, the EE-MB methods are designed in such a way as to allow the straightforward introduction of periodic boundary conditions, so that they give a promising alternative to current simulation techniques for molecular liquids.

An important advantage of the EE-MB methods is that they can easily be employed with any electronic structure package that allows for using a field of background point charges. Furthermore once implemented for a given electronic structure package, the EE-MB program is available for all electronic structure levels available in that package. It is very efficient for any density functional or wave function theory that has analytic gradients, and it can provide a substantial savings in cost for large systems.

ACKNOWLEDGMENT.

This work was supported in part by the National Science Foundation under grant nos. CHE07-04974 and ITR-0428774.

Supporting Information Available: Geometry and B3LYP/6-31+G(d,p) energy and gradient for the 64-molecule cluster. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

(1) Alder, T. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.

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

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

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

(5) Darden, T. A. In *Computational Biochemistry and Biophysics*; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Wantanbe, M., Eds.; Marcel Dekker Inc.: New York, 2001; pp 91−115.

(6) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation: From Algorithms to Applications;* Academic Press: New York, 2002; pp 32−40.

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

(8) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35. Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

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

(10) Raghavachari, K.; Anderson, J. B. *J. Phys. Chem.* **1996**, *100*, 12960.

(11) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, 1133. Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(12) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *J. Phys. Chem. A* **2006**, *110*, 640.

(13) McGrath, M. J.; Siepmann, I. J.; Kuo, I.-F. W.; Mundy, C. J. *Mol. Phys.* **2006**, *104*, 3619.

(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.; Uebayashi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.

(16) Komeiji, Y.; Nakano, T.; Fukuwaza, K.; Ueno, Y.; Inadomi, Y.; Nemoto, T.; Uebayasi, M.; Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2003**, *372*, 342.

(17) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832. Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129.

(18) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.

(19) Kulkarni, A.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2004**, *121*, 5043.

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

(21) Christie, R. A.; Jordan, K. D. *Struct. Bonding (Berlin)* **2005**, *116*, 27.

(22) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2006**, *433*, 182.

(23) Fedorov, D. G.; Kitaura, K. In *Modern Methods for Theoretical Physical Chemistry of Biopolymers*; Starikov, E. B., Lewis, J. P., Tanaka, S., Eds.; Elsevier: Amsterdam, 2006; pp 3−38.

(24) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagase, S. *J. Comput. Chem.* **2007**, *28*, 1476.

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

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

(27) Fedorov, D. G.; Ishida, T.; Uebayasi, M.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 2722.

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

(29) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput*. Accepted for pubication.

(30) Amovilli, C.; Cacelli, I.; Campanile, S.; Prampolini, G. *J. Chem. Phys.* **2002**, *117*, 3003.

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

(32) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846.

(33) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Comput. Chem.* **2004**, *25*, 1431.

(34) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145.

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

(36) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

(37) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102. Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.

(38) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777. Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2007**, *111*, 5111.

(39) Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 24104.

(40) Antes, I.; Thiel, W. *ACS Symp. Ser.* **1998**, *712*, 50.

(41) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.

(42) Amara, P.; Field, M. *J. Theor. Chem. Acc.* **2003**, *109*, 43.

(43) Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 3991.

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

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

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

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

(48) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. In *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; p 576.

(49) Lenosky, T. J.; Kress, J. D.; Kwon, I.; Voter, A. F.; Edwards, B.; Richards, D. F.; Yang, S.; Adams, J. B. *Phys. Rev. B* **1997**, *55*, 1528.

(50) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Robb, G. E. S. 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. *2003, Gaussian03-version c01*; Gaussian Inc.: Wallingford, CT, 2004.

(51) Dahlke, E. E.; Truhlar, D. G. *MBPAC 2007-2*; University of Minnesota: Minneapolis, MN, 2007.

(52) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2006**, *2*, 1370.

(53) Attard, P. *Phys. Rev. A* **1992**, *45*, 5649.

(54) Sauer, J.; Sierka, M. *J. Comput. Chem.* **2000**, *16*, 1470.

(55) Bühl, M.; Mauschich, F. T. *PhysChemChemPhys* **2002**, *433*, 5508.

(56) Clark, L. A.; Sierka, M.; Sauer, J. *Stud. Surf. Sci. Catal.* **2002**, *142 A*, 643.

(57) Roca, M.; Martí, S.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J.; Williams, I. H. *J. Am. Chem. Soc.* **2003**, *125*, 7726.

(58) Rega, N.; Iyengar, S. S.; Voth, G. A.; Schlegel, H. B.; Vreven, T.; Frisch, M. J. *J. Phys. Chem. B* **2004**, *108*, 4210.

(59) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *157*, 227.

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

(61) Théry, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. C. *J. Comput. Chem.* **1994**, *15*, 269.

(62) Thompson, M. A.; Glendening, E. D.; Feller, D. *J. Phys. Chem.* **1994**, *98*, 10465.

(63) Stanton, R. V.; Hartsough, D. S.; Merz, K. M. *J. Comput. Chem.* **1995**, *16*, 113.

(64) Bakowies, D.; Thiel, W. *J. Phys. Chem. A* **1996**, *100*, 10580.

(65) Ferré, N.; Assfeld, X.; Rivail, J.-L. *J. Comput. Chem.* **2002**, *23*, 610.

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

(67) Pu, J.; Gao, J.; Truhlar, D. G. *ChemPhysChem* **2005**, *69*, 1853.

(68) Laasonen, K.; Car, R.; Lee, C.; Vanderbuilt, D. *Phys. Rev. B* **1991**, *43*, 6796.

(69) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*, 17953.

(70) Stadler, R.; Wolf, W.; Pudloucky, R.; Kresse, G. *Phys. Rev. B* **1996**, *54*, 1729.

(71) Stokbro, K. *Phys. Rev. B* **1996**, *53*, 6869.

(72) Moroni, E. G.; Kresse, G.; Hafner, J.; Furthmüller, J. *Phys. Rev. B* **1997**, *56*, 15629.

(73) Pulay, P.; Saebo, S.; Malagoli, M.; Baker, J. *J. Comput. Chem.* **2005**, *26*, 599.

(74) Paier, J.; Hischi, R.; Marsman, M.; Kresse, G. *J. Chem. Phys*. **2005**, *122*, 234102.

(75) Dovesi, R.; Civalleri, B.; Orlando, R.; Roetti, C.; Saunders, V. R. *Rev. Comp. Chem.* **2005**, *21*, 1.

(76) Hamaan, D. R.; Schlüter, M.; Chiang, C. *Phys. Rev. Lett.* **1979**, *43*, 1494.

(77) Pacios, L. F.; Christiansen, P. A. *J. Chem. Phys.* **1985**, *82*, 2004.

(78) Fattebert, J.-L.; Gygi, F. *Phys. Rev. B* **2006**, *73*, 115124.

(79) Ianuzzi, M.; Kirchner, B.; Hutter, J. *Chem. Phys. Lett.* **2006**, *421*, 16.

CT700223R

JCTC Journal of Chemical Theory and Computation

Accurate Induction Energies for Small Organic Molecules: 1. Theory

Alston J. Misquitta[†,‡] and Anthony J. Stone*[,†]

University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, U.K., and
University College London, 20 Gordon Street, London WC1H 0AJ, U.K.

Received May 3, 2007

Abstract: The induction energy often plays a very important role in determining the structure and properties of clusters of organic molecules, but only in recent years has an effort been made to include this energy in such calculations, notably in the field of organic crystal structure prediction. In this paper and the following one in this issue we provide ab initio methods suitable for the accurate inclusion of the induction energy for molecules containing as many as 30 atoms or so. These techniques are based on Symmetry-Adapted Perturbation Theory using Density Functional Theory [SAPT(DFT)] and use distributed polarizabilities computed using the recently developed density-fitting algorithm with constrained refinement. With this approach we are able to obtain induction models of varying complexity and study the effects of overlap and related numerical issues. Basis set effects on the exact and asymptotic induction energies are investigated, and the roles of higher-order induction energies and many-body effects are explored.

I. Introduction

The induction energy plays an important role in determining the structures of clusters of polar molecules. The cooperative nature of the induction means that, for polar molecules, induction effects dominate the many-body contributions to the interaction energy. These many-body effects can be very important in determining the structures of clusters of molecules. For example, three- and four-body effects have been shown to be responsible for the tetrahedral structure of liquid water.[1,2] However, this interaction energy component is often neglected or treated incorrectly. This is not only because it is hard to calculate accurately but also because important aspects of the induction energy are still poorly understood.

There are many features that make the induction energy hard to handle. First, it is not pair-additive. The induction energy of a particular molecule, $-(1/2)\alpha F^2$ in its simplest form, depends on the square of the total electric field F due to its neighbors, and the fields of different neighbors may interfere constructively or destructively.[3] Second, it is cooperative: the charge distribution of each molecule is polarized by the electric field of its neighbors, and it is the modified charge distribution of each molecule that is the source of the field that polarizes the others. This cooperative behavior is important in clusters and condensed phases of polar molecules and favors the hydrogen-bonded networks that are seen for example in water. Since each molecule is polarized by all the others, it is necessary to solve coupled equations for the modified charge distributions. This can usually be carried out by a simple iterative procedure, but it is a time-consuming additional step in a simulation.

The most general way to calculate the induction[4] uses the frequency-dependent density susceptibility,[5] or FDDS, $\alpha(\mathbf{r},\mathbf{r}'|\omega)$, which describes the change in charge density at $\mathbf{r}$ due to a delta-function change in electrostatic potential at $\mathbf{r}'$ oscillating at frequency ω. To describe induction we need only the static FDDS, $\alpha(\mathbf{r},\mathbf{r}'|0)$, and this can be calculated efficiently and accurately by modern methods. The resulting description is however much too cumbersome for practical use; it would be necessary to solve a set of coupled integral equations for the changes to the electron density of each molecule and then to carry out integrals over each molecule to determine its induction energy. Consequently we need to

* Corresponding author phone: +44 1223 336375; fax: +44 1223 336362; e-mail: ajs1@cam.ac.uk.

[†] University Chemical Laboratory.

[‡] University College London.

10.1021/ct700104t CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/10/2007

extract a description in terms of polarizabilities. For all but the smallest of molecules, single-site descriptions, using the overall molecular polarizabilities, are inadequate, and the polarizability needs to be described in a distributed way, as the charge distribution does.

This does not however solve all the problems. At short range, when the molecular charge densities overlap, the distributed-polarizability description is subject to penetration error, just like the distributed-multipole description of the electrostatic interaction. Using the FDDS we could describe the penetration effects correctly, but unfortunately that is usually impracticable.

This is by no means the end of the story. The polarizability, even in the accurate form of the FDDS, describes the *linear* response of the molecule to external fields and gives the induction energy only to second order in perturbation theory.[4] Moreover, as usually formulated, it ignores effects arising from electron exchange between molecules at short range. For a more complete description, we need to include higher-order terms in the perturbation series and to include the effects of exchange.[6,7]

In fact a large part of the higher-order effects for clusters can be recovered if the polarization problem is solved self-consistently,[3] by the iterative procedure mentioned above, but this still treats the response of each individual molecule as linear in the field. Higher-order effects, described by hyperpolarizabilities, are not included in such a treatment and become increasingly significant at short distances. As we shall see in this paper, these neglected effects can result in a significant error in the dimer energy and geometry.

As for the exchange effects, it has become clear from recent calculations using Symmetry-Adapted Perturbation Theory (SAPT) that they make a very significant contribution to the total induction energy.[7,8]

In this paper and the following one in this issue[1] we attempt to obtain a practical procedure for calculating accurate induction energies for assemblies of molecules in clusters or in the condensed phase. Here we discuss the theoretical issues associated with accurate calculations of the nonexpanded and expanded induction energies. The numerical issues associated with basis sets and model building will be discussed in part 2.

II. General Overview

For the interaction between two molecules, the second-order induction energy can be accurately computed using symmetry-adapted perturbation theory (SAPT)[7] or the more recent version of SAPT based on a density-functional treatment of the monomers (SAPT(DFT)[9−12] or the very similar and independently derived DFT-SAPT[13−16]). In the following we refer to SAPT(DFT) for brevity, but it should be understood that the DFT-SAPT method is essentially the same. The superior computational scalings and accuracies of the SAPT(DFT) expressions make this theory the method of choice, particularly for organic molecules, for which the SAPT expressions are usually too expensive computationally to be evaluated. The SAPT(DFT) expression for the second-order induction involves the frequency-dependent density susceptibility (FDDS) at zero frequency,[11] which can be evaluated quite efficiently using coupled Kohn−Sham (CKS) theory (also known as Kohn−Sham linear response theory). This expression does not include exchange effects; the second-order exchange-induction energy cannot be written in terms of the FDDS and is calculated using scaling rules[11,12] which have been demonstrated to result in rather accurate energies.

For polar molecules with large polarizabilities, the higher-order induction and exchange-induction energies (in which we include terms of third order and above) can make significant contributions to the two-body interaction energy. (We include in the category of polar molecules any molecule that gives rise to large electric fields in its neighborhood, whether or not it has a significant dipole moment. The interaction between such molecules is dominated by the electrostatic energy.) These higher-order effects are strongest in hydrogen-bonded complexes, where they can account for as much as 10−15% of the total two-body interaction energy. In most SAPT and SAPT(DFT) calculations of the intermolecular energy the higher-order energies have been estimated using the so-called $\delta^{\rm HF}_{\rm int,resp}$ term, defined as the difference between the supermolecular Hartree−Fock interaction energy of the dimer and certain low-order SAPT energy terms.[7,17,18] This procedure has been shown to result in interaction energy potentials of high accuracy for hydrogen-bonded complexes like the water dimer,[2,19] but recent evidence seems to suggest that the $\delta^{\rm HF}_{\rm int,resp}$ term may not be suitable for non-hydrogen-bonded complexes.[20] We will return to this issue below. In any case, the $\delta^{\rm HF}_{\rm int,resp}$ term is cumbersome to calculate as it requires a supermolecular Hartree−Fock calculation in the dimer basis (so as to avoid the basis-set superposition error) and a low-order SAPT calculation, in addition to the SAPT(DFT) calculation. Consequently an alternative means of estimating the higher-order contributions to the interaction energy is needed.

Recently, the third-order SAPT interaction energy components have been derived and implemented, though without the inclusion of intramonomer correlation effects.[20] It has often been assumed that the third-order energies would account for most of the higher-order contributions to the interaction energy. However, Patkowski et al.[20] have demonstrated that while this is true for non-hydrogen-bonded complexes, the third-order terms may account for less than half of the higher-order energies for hydrogen-bonded complexes. It is believed that this is the case for two reasons: first, the higher-order energies are dominated by induction and exchange-induction effects,[20] and second, the induction series is known to be divergent due to the presence of Coulomb singularities in the interaction operator.[21]

In spite of the problems associated with the higher-order interaction energy components, we shall demonstrate that the third-order induction and exchange-induction energies can form a good and computationally convenient approximation to the true higher-order energies if evaluated within the SAPT(KS) theory,[9,11] that is, using Kohn−Sham orbital energies and eigenvalues. This procedure has many advantages: (1) The higher-order energies for non-hydrogen-bonded complexes are recovered very accurately. (2) While there will be non-negligible errors made in hydrogen-bonded geometries, these will be much smaller than the errors incurred if the higher-order corrections were ignored altogether. (3) The interaction energies are obtained in one calculation. (4) And finally, the third-order induction and exchange-induction energies are the least computationally demanding of the third-order energy components and so do not add significantly to the overall computational cost of the SAPT(DFT) method.

The many-body contribution to the interaction energy can be very important for polar clusters, in which many-body effects can account for as much as 15−30% of the interaction energy.[22,23] This is not surprising, as the induction energy, which is very important for such systems, is strongly nonadditive in nature.[3] The dominant contribution to the many-body energy arises from the three-body energy which can be computed using the three-body formulation of SAPT (see ref 7 for a review) or using supermolecular techniques,[22,24] but the computational expense is so large as to make these methods applicable to systems of a few atoms only. Fortunately, for polar systems, which include most organic molecules, the many-body contributions are dominated by many-body induction effects which can be well approximated using damped classical polarizable models if accurate molecular polarizabilities and multipole moments are known.

This paper is organized as follows: In section IV.1 we outline the theoretical details of the SAPT(DFT) expression for the second-order induction energy. In subsection IV.1.2 we explore and assess ways of including the higher-order induction and exchange-induction energies using several examples. In section IV.2 we briefly describe ways of including the many-body contributions to the induction energy using the damped classical polarizable model. The damped classical polarizable model is also used to calculate the asymptotic induction energies. This needs the molecular polarizabilities in distributed form. In section IV.3 we describe the distribution method based on the constrained density-fitting procedure and present a method for optimizing the resulting distributed polarizabilities. In section V we conclude with a summary of the main results of this paper.

III. Notation

If electron exchange between molecules is ignored, which is a good approximation at large separations, the interaction energy can be obtained using standard perturbation theory. This is conventionally described as the polarization approximation, though this is an unsatisfactory terminology, particularly in the context of the induction energy. The contribution that is usually called the induction energy, and denoted E_{ind}, appears at second order, but we denote it here as $E^{(2)}_{\mathrm{ind,pol}}$ to distinguish it explicitly from the exchange-induction $E^{(2)}_{\mathrm{ind,exch}}$. $E^{(2)}_{\mathrm{ind,pol}}$ is the term defined as the induction and denoted E_{ind}, in SAPT and SAPT(DFT).[12] However there are also induction energy contributions at higher orders, $E^{(n)}_{\mathrm{ind,pol}}$. The second-order term $E^{(2)}_{\mathrm{ind,pol}}$ can be expressed in a ‘nonexpanded’ form which remains valid at any intermolecular separation, however small, but it is conventionally expressed as a power series in $1/R$. This power-series form is often referred to as the classical model, and we denote the damped version of this model by $E^{(2)}_{\mathrm{ind,d-class}}$. Higher-order contributions can be expanded in the same way. We should have $E^{(2)}_{\mathrm{ind,pol}} \sim E^{(2)}_{\mathrm{ind,d-class}}$ for medium to large separations, but it turns out that the basis-set converged value of $E^{(2)}_{\mathrm{ind,pol}}$ can be an order of magnitude larger (i.e., more negative) than $E^{(2)}_{\mathrm{ind,d-class}}$ at equilibrium geometry and even larger at shorter distances. It has recently been shown[21] that $E^{(2)}_{\mathrm{ind,pol}}$ is too large in magnitude because of the Coulomb singularities in the interaction operator. These singularities are absent in the expanded form of the operator, so it is not surprising that $E^{(2)}_{\mathrm{ind,d-class}}$ is much smaller in magnitude than $E^{(2)}_{\mathrm{ind,pol}}$. The Coulomb singularities also result in a very large exchange-induction energy, $E^{(2)}_{\mathrm{ind,exch}}$, which is positive and significantly quenches $E^{(2)}_{\mathrm{ind,pol}}$. Therefore neither energy is meaningful on its own.

Because of these complications, we believe that rather than referring to the conventional definitions[12] of the induction and exchange-induction energies separately, it is much better to define the induction energy as the sum. That is, the nth order induction energy is

$$E^{(n)}_{\mathrm{ind,tot}} = E^{(n)}_{\mathrm{ind,pol}} + E^{(n)}_{\mathrm{ind,exch}} \tag{1}$$

Since $E^{(n)}_{\mathrm{ind,exch}}$ decays exponentially with increasing R, $E^{(n)}_{\mathrm{ind,tot}}$ and $E^{(n)}_{\mathrm{ind,pol}}$ both tend to $E^{(n)}_{\mathrm{ind,d-class}}$ asymptotically. However, as will be demonstrated in sections V and VI of part 2, $E^{(n)}_{\mathrm{ind,tot}}$ agrees far better with $E^{(n)}_{\mathrm{ind,d-class}}$ at all distances than $E^{(n)}_{\mathrm{ind,pol}}$ does.

The total interaction energy including terms up to order n is denoted by $U^{(n)}$ rather than the more conventional $E^{(n)}_{\mathrm{int}}$. This has been done so as to avoid possible confusion arising from the similarity of the subscripts ‘ind’ and ‘int’.

IV. Theory

IV.1. Induction Contributions to the Two-Body Energy.

In the two-body energy, the induction contributes to terms of second and higher orders in the interaction operator. The second-order induction is the most important, constituting between 85% and 96% of the total two-body induction energy.

IV.1.1. At Second Order: $E^{(2)}_{ind,pol}$ and $E^{(2)}_{ind,exch}$. From the polarization expansion,[6,7] $E^{(2)}_{\mathrm{ind,pol}}$ for molecule X is

$$E^{(2)}_{\mathrm{ind,pol}}(X) = \sum_{r\neq 0} \frac{|\langle \Phi^X_0|\hat{V}|\Phi^X_r\rangle|^2}{E^X_0 - E^X_r} \tag{2}$$

where Φ^X_r and E^X_r are the eigenstates and energy eigenvalues of the monomer Hamiltonian H_X, and $\hat{V}$ is the perturbation due to the electrostatic potential arising from the rest of the system. $E^{(2)}_{\mathrm{ind,pol}}(X)$ can be interpreted as the second-order response of monomer X to the static field $\hat{V}$. One can show that eq 2 can be rewritten in terms of the frequency-dependent density susceptibility (FDDS) $\alpha_X(\mathbf{r},\mathbf{r}'|\omega)$ evaluated at zero frequency[4]

$$E^{(2)}_{\mathrm{ind,pol}}(X) = -\frac{1}{2}\int\int \alpha_X(\mathbf{r},\mathbf{r}'|0)V(\mathbf{r})V(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{3}$$

where

$$\alpha_X(\mathbf{r}, \mathbf{r}'|\omega) = 2\sum_{r\neq 0}\frac{E_r^X - E_0^X}{(E_r^X - E_0^X)^2 - \omega^2}\langle\Phi_0^X|\hat{\rho}_X(\mathbf{r})|\Phi_r^X\rangle\langle\Phi_r^X|\hat{\rho}_X(\mathbf{r}')|\Phi_0^X\rangle \quad (4)$$

and $V = \int\rho_Y^{\mathrm{tot}}(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'|\mathrm{d}\mathbf{r}'$ is the electrostatic potential of the rest of the system. In the equation above, $\hat{\rho}_X(\mathbf{r}) = -\sum_{i\in X}\delta(\mathbf{r} - \mathbf{r}_i)$ is the electron density operator. The FDDS describes the linear response of the electron density to a frequency-dependent perturbation.

To calculate $E^{(2)}_{\mathrm{ind,pol}}$ for an interacting pair of molecules within SAPT(DFT), the electrostatic potentials of the unperturbed monomers are evaluated using the Kohn−Sham orbitals, and the FDDS is evaluated using coupled Kohn−Sham theory (CKS) [25−27] (also known as linear-response DFT). In CKS theory, the FDDS takes the form

$$\alpha(\mathbf{r},\mathbf{r}'|\omega) = \sum_{iv,i'v'} C_{iv,i'v'}(\omega)\phi_i(\mathbf{r})\phi_v(\mathbf{r})\phi_{i'}(\mathbf{r}')\phi_{v'}(\mathbf{r}') \quad (5)$$

where the subscripts i and i' (v and v') denote occupied (virtual) molecular orbitals, and ϕ_i is a molecular orbital. In CKS theory (and coupled Hartree−Fock theory (CHF)) the coefficients $C_{iv,i'v'}(\omega)$ can be written as[26]

$$C_{iv,i'v'}(\omega) = 4[(\mathbf{H}^{(2)}\mathbf{H}^{(1)} - \hbar^2\omega^2\mathbf{I})^{-1}\mathbf{H}^{(2)}]_{iv,i'v'} \quad (6)$$

where $\mathbf{I}$ is the unit matrix, and the $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ matrices, called the electric and magnetic Hessians, respectively, are defined in the CKS theory (in the adiabatic approximation[25,26]) as follows

$$\mathbf{H}^{(1)}_{iv,i'v'} = (e_v - e_i)\delta_{iv,i'v'} + 4(iv|i'v') - c_x[(ii'|vv') + (iv'|i'v)] + 4\int\phi_i\phi_v\phi_{i'}\phi_{v'}\frac{\delta(v_{\mathrm{xc}} - c_x v_{\mathrm{x}})}{\delta\rho}\mathrm{d}^3\mathbf{r} \quad (7)$$

and

$$\mathbf{H}^{(2)}_{iv,i'v'} = (e_v - e_i)\delta_{iv,i'v'} - c_x[(ii'|vv') - (iv'|i'v)], \quad (8)$$

where e_i is the Kohn−Sham energy eigenvalue of molecular orbital ϕ_i, c_x is the fraction of the Hartree−Fock exchange included in the exchange-correlation (XC) functional ($c_x = 0$ for a nonhybrid functional), v_{x} is the exchange part of v_{xc}, and the two-electron coulomb integral $(ij|kl) = \langle\phi_i(1)\phi_j(1)|1/r_{12}|\phi_k(2)\phi_l(2)\rangle$. CKS theory is, in principle, exact if the exact exchange-correlation functional is used. In practice, only approximations are known, and from extensive numerical experiments it has been concluded[11,12,28] that the asymptotically corrected[29,30] PBE0[31] exchange-correlation functional[12,32] is the most suitable for accurate interaction energies. For large molecules however, it is too expensive computationally to evaluate the last integral in eq 7 using v_{xc} and v_{x} from the PBE functional. A more practical approach is to use the exchange-only LDA functional in the last term of eq 7. This approximation results in a small (less than 1%) loss in accuracy which is more than compensated by an order of magnitude reduction in computational expense.[12]

In order to implement the asymptotic correction, accurate vertical ionization potentials (IPs) are needed for the monomers. When they are not available experimentally, good estimates may be obtained from the difference between the energies of the N and $N - 1$ electron systems. The PBE0 functional is best suited for this calculation too as tests on atoms, diatoms, and small organic molecules have shown that it gives IPs with mean errors centered about 0.0 au with a standard deviation of only 0.007 au.[33]

As with the dispersion energy,[10,12] density-fitting techniques can be used to make the evaluation of the $E^{(2)}_{\mathrm{ind,pol}}$ more efficient. Using density-fitting, the molecular orbital products $\phi_i(\mathbf{r})\phi_v(\mathbf{r})$ that appear in eq 5 can be expanded as

$$\phi_i(\mathbf{r})\phi_v(\mathbf{r}) = \sum_p D_{iv,p}\chi_p(\mathbf{r}) \quad (9)$$

where $\{\chi\}$ is an auxiliary basis set, and the coefficients $D_{iv,p}$ are determined by least-squares. The density-fitted FDDS takes the form

$$\alpha(\mathbf{r},\mathbf{r}'|\omega) = \sum_{p,q}\tilde{C}_{pq}(\omega)\chi_p(\mathbf{r})\chi_q(\mathbf{r}') \quad (10)$$

where the $\tilde{C}_{pq}(\omega)$ are the transformed coefficients given by $\tilde{C}_{pq}(\omega) = \sum_{iv,i'v'}D_{iv,p}C_{iv,i'v'}(\omega)D_{i'v',q}$. Within Kohn−Sham theory, the total charge density ρ_Y^{tot} of closed-shell systems is given by

$$\rho_Y^{\mathrm{tot}}(\mathbf{r}) = \sum_\beta Z_\beta\delta(\mathbf{r} - \mathbf{R}_\beta) - 2\sum_j|\phi_j(\mathbf{r})|^2 \quad (11)$$

where Z_β and $\mathbf{R}_\beta$ are the nuclear charge and position, and j labels the occupied orbitals. Consequently the electrostatic potential V of the rest of the system can be written as

$$V(\mathbf{r}) = \sum_\beta\frac{Z_\beta}{|\mathbf{r} - \mathbf{R}_\beta|} - 2\sum_j\int\frac{|\phi_j(\mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|}\mathrm{d}\mathbf{r}' \quad (12)$$

where the sums run over the nuclei and orbitals of the rest of the system, which in this case is just monomer Y. Using the density-fitted FDDS and the above expression for V in eq 3 we obtain the density-fitted form of $E^{(2)}_{\mathrm{ind,pol}}$ for monomer X

$$E^{(2)}_{\mathrm{ind,pol}}(X) = -\frac{1}{2}\sum_{pq}M_p^Y\tilde{C}^X_{pq}(0)M_q^Y \quad (13)$$

where M_p^Y is defined as

$$M_p^Y = L_p^Y - 2\sum_{q'}\sum_j J_{pq'}D^Y_{jj,q'} \quad (14)$$

and we have used the definitions

$$J_{pq'} = \int\int\frac{\chi_p(\mathbf{r})\chi_{q'}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \quad (15)$$

and

$$L_p^Y = \sum_\beta\int\frac{Z_\beta}{|\mathbf{r} - \mathbf{R}_\beta|}\chi_p(\mathbf{r})\mathrm{d}\mathbf{r} \quad (16)$$

The evaluation of $E^{(2)}_{\mathrm{ind,pol}}$ using eq 13 involves a computational cost that scales as $O(m^2)$, where m is the size of the auxiliary basis. This is smaller than the $O(n_o^2 n_v^2)$ without density-fitting, but, much more importantly, the computationally demanding 4-center 2-electron integral transformation is avoided in the evaluation of eq 13, having been replaced by the 2-center 2-electron integrals $J_{pq'}$.

Futhermore, when monomer basis sets are used, only J_{pq} and L^Y_p need to be recomputed for each dimer geometry. This can result in a considerable savings in computational effort as only $O(m^2)$ operations are needed to evaluate the induction energy at each dimer geometry. However, see section VI in part 2 for some of the numerical issues associated with using such basis sets.

In part 2 we will need to evaluate the second-order induction energy of a molecule and a point charge. For such a system, the electrostatic potential is $V = Q/|\mathbf{r} - \mathbf{R}_Q|$ where $\mathbf{R}_Q$ is the location of the point charge and Q its value. The resulting induction energy is simply

$$E^{(2)}_{\mathrm{ind,pol}} = -\frac{1}{2}\sum_{pq} L^Q_p \tilde{C}_{pq}(0) L^Q_q \qquad (17)$$

where $L^Q_p = \int [Q/|\mathbf{r} - \mathbf{R}_Q|]\chi_p(\mathbf{r})\mathrm{d}\mathbf{r}$.

The exchange-induction energy at second order, $E^{(2)}_{\mathrm{ind,exch}}$, quenches $E^{(2)}_{\mathrm{ind,pol}}$ significantly. $E^{(2)}_{\mathrm{ind,exch}}$ cannot be expressed in terms of the FDDS and electron densities of the monomers, so it is estimated from the SAPT(KS) energies (denoted by ‘KS’) using a scaling relation[12]

$$E^{(2)}_{\mathrm{ind,exch}} \approx E^{(2)}_{\mathrm{ind,exch}}(\mathrm{KS}) \times \frac{E^{(2)}_{\mathrm{ind,pol}}}{E^{(2)}_{\mathrm{ind,pol}}(\mathrm{KS})} \qquad (18)$$

The large quenching of $E^{(2)}_{\mathrm{ind,pol}}$ by $E^{(2)}_{\mathrm{ind,exch}}$ is believed to be an effect of excessive electron tunneling due to the electron−nuclear Coulomb singularities in the interaction operator.[21] These singularities are also responsible for the divergence of the perturbation theory. It has been shown that a convergent perturbation theory can be built using a regularized form of the interaction operator, that is, one in which the singularities arising from the electron−nuclear terms are removed.[21] It is also possible that for a regularized version of SAPT(DFT), the exchange-induction terms would not be needed. Preliminary evidence from our group suggests that this may well be the case.[34]

In summary, then, we recommend that the second-order induction energy be expressed according to eq 1, that is, as

$$E^{(2)}_{\mathrm{ind,tot}} = E^{(2)}_{\mathrm{ind,pol}} + E^{(2)}_{\mathrm{ind,exch}} \qquad (19)$$

with $E^{(2)}_{\mathrm{ind,pol}}$ calculated by coupled Kohn−Sham theory as described above, and $E^{(2)}_{\mathrm{ind,exch}}$ given by eq 18.

IV.1.2. Higher-Order Two-Body Energies. Contributions to the two-body interaction energy from terms beyond the second order in perturbation theory, denoted by $U^{(3-\infty)}$, are often large and cannot be neglected. For polar molecules like water, induction energies dominate these higher-order terms, which can constitute as much as 15% of the equilibrium binding energy of the dimer. On the other hand, for nonpolar molecules, the higher-order energies contribute only about 3−5% of the total interaction energy and could even be ignored if high accuracies are not needed.

$U^{(3-\infty)}$ is often approximated by the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ correction, defined as[17,18]

$$\delta^{\mathrm{HF}}_{\mathrm{int,resp}} = U^{\mathrm{HF}} - (E^{(10)}_{\mathrm{elst}} + E^{(10)}_{\mathrm{exch}} + E^{(20)}_{\mathrm{ind,tot,resp}}) \qquad (20)$$

where $E^{(10)}_{\mathrm{elst}}$, $E^{(10)}_{\mathrm{exch}}$, and $E^{(20)}_{\mathrm{ind,tot,resp}}$ are the SAPT corrections with no intramonomer correlation effects included (The subscript ‘resp’ indicates that the induction energy is calculated with response effects included.[7]), and U^{HF} is the supermolecule Hartree−Fock interaction energy computed with the counterpoise correction. Bear in mind that $E^{(20)}_{\mathrm{ind,tot,resp}}$ has been defined using eq 1 and includes the exchange-induction contribution. The δ^{HF} term approximates the third- and higher-order contributions to the interaction energy but must be defined within SAPT as it has no counterpart in SAPT(DFT). Unfortunately, this term is too cumbersome to calculate on a routine basis, as its evaluation using eq 20 involves a supermolecule Hartree−Fock calculation in the dimer basis in addition to a low-order SAPT calculation. Furthermore, there is evidence that it may be a poor approximation to the higher-order energies for nonpolar systems[20] (but see the discussion in section V).

Another means of approximating $U^{(3-\infty)}$ is through the damped classical polarizable model. The derivation of the induction energy in a classical polarizable model is given in ref 3. Here we reproduce the final expressions in their general form. The damped classical induction energy of molecule A in a cluster is

$$E_{\mathrm{ind,d-class}}(A) = \frac{1}{2}\sum_{a\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{tu} \Delta Q^a_t f_{(tu)}(\beta R_{ab}) T^{ab}_{tu} Q^b_u \qquad (21)$$

where Q^a_t is the multipole moment operator for moment t at site a, and T^{ab}_{tu} is the interaction tensor[3] which describes the interaction between a multipole Q_u at b and a multipole Q_t at a. $f_{(tu)}(\beta R_{ab})$ is a damping function, which is conventionally assumed, in the absence of evidence to the contrary, to depend only on the distance R_{ab} between sites a and b and not on their relative orientation; the parameter β specifies the strength of the damping and may depend on the nature of the sites. ΔQ^a_t is the change in multipole moment t at a due to the self-consistent polarization of site a in the field of all sites on other molecules and is given by

$$\Delta Q_t^{\,a} = -\sum_{a'\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{t'v} \alpha^{aa'}_{tt'} f_{(t'v)}(\beta R_{a'b}) T^{a'b}_{t'v}(Q^b_v + \Delta Q^b_v) \qquad (22)$$

where $\alpha^{aa'}_{tt'}$ is the distributed polarizability for sites (a, a') which describes the response of the multipole moment component Q^a_t at site a to the t'-component of the field at site a'. Notice that eq 22 must be solved iteratively for all molecules in the system, as the ΔQ occur on both sides of the equation. In the case of a dimer the sum over B just includes the other member of the dimer. If ΔQ is omitted from the right-hand side, we recover the damped classical approximation of the second-order induction energy, $E^{(2)}_{\mathrm{ind,d-class}}$. At the mth iteration, we additionally obtain the

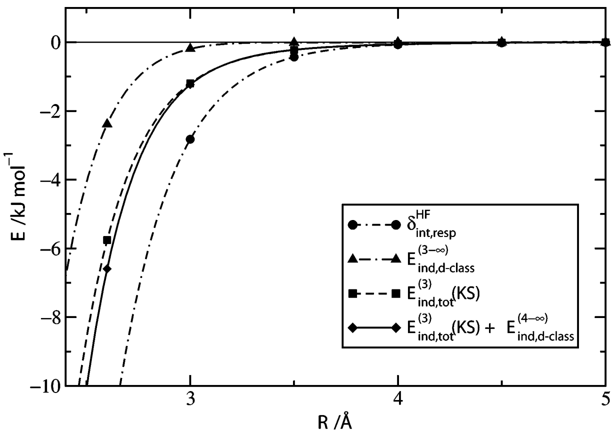


Figure 1. Approximations to the higher-order induction and exchange-induction energies for the water dimer. The relative orientations of the water molecules are fixed at their minimum geometry,[35] and the center-of-mass separation is varied. The minimum is located close to $R = 3.0$ Å. The classical models used a damping coefficient $\beta = 1.93$ au (see section V in part 2).

damped classical approximation to $E^{(m+2)}_{\text{ind,tot}}$. Call this approximation $E^{(m+2)}_{\text{ind,d}-\text{class}}$. The quantum expressions for the higher-order induction energies involve, besides the FDDS, which is a linear response function, the quadratic and higher-order response functions. The classical polarization model completely neglects these higher-order response functions. Furthermore, orbital overlap effects are neglected by the classical model, and these effects become increasingly important with increasing order in perturbation theory. The incorporation of the damping function attempts to correct for this neglect, but little is known about the form that it should take. Therefore, even though the damped classical polarizable model works reasonably well for the second-order induction energy, it is not reliable for the higher-order two-body induction effects.

The failure of the classical polarization models to recover the higher-order energies is clearly illustrated in Figure 1 for the water dimer. For this system, comparisons with interaction energies calculated using CCSD(T) suggest that that $\delta^{\text{HF}}_{\text{int,resp}}$ is a good estimate of the higher-order energies. The damped classical estimate of these energies, $E^{(3-\infty)}_{\text{ind,d}-\text{class}}$, is clearly inadequate, being an order of magnitude smaller than the $\delta^{\text{HF}}_{\text{int,resp}}$ correction for energetically relevant center-of-mass separations.

An alternative to the above methods is to approximate $U^{(3-\infty)}$ by the third-order induction energy and include the missing terms of fourth and higher orders using the damped classical polarizable model, that is

$$U^{(3-\infty)} \approx E^{(3)}_{\text{ind,tot}} + E^{(4-\infty)}_{\text{ind,d}-\text{class}} \qquad (23)$$

Here $E^{(4-\infty)}_{\text{ind,d}-\text{class}}$ is the damped classical induction energy summed from the second iteration onward (i.e., to convergence). The third-order energies are expected to dominate $U^{(3-\infty)}$, so this approximation would include most of the overlap effects that are missing from the classical polarizable model. The SAPT expressions for $E^{(3)}_{\text{ind,pol}}$ and $E^{(3)}_{\text{ind,exch}}$ have been derived[20,38] but without the inclusion of orbital relaxation effects. That is, when used with Kohn−Sham orbitals and eigenvalues, these energies are obtained at the SAPT-(KS) level of theory. We will denote the resulting approximation by $U^{(3-\infty)}$(KS) which, using eq 1, is defined as

$$U^{(3-\infty)}(\text{KS}) = E^{(3)}_{\text{ind,tot}}(\text{KS}) + E^{(4-\infty)}_{\text{ind,d}-\text{class}} \qquad (24)$$

Orbital relaxation effects have been demonstrated to play a relatively minor role for the second-order induction.[11] In ref 11 this was argued to be at least in part due to a cancellation of errors. As there is no reason to expect the same to happen for the third-order energies, one might question the use of the SAPT(KS) expressions here. However, $U^{(3-\infty)}$ is relatively small in magnitude compared with the second-order energy, so any errors incurred by the use of the SAPT(KS) expressions are probably less important.

From Figure 1 we see that $U^{(3-\infty)}$(KS) is a far better approximation to the higher-order energy contributions than the damped classical polarizable model alone, but the higher-order terms are probably still underestimated by this approximation. In Table 1 we report second-order interaction energies and the various estimates of the third-order terms discussed above for a few dimers. For the water dimer, taking $U^{\text{CCSD(T)}}$ as a reference, we see that while adding $\delta^{\text{HF}}_{\text{int,resp}}$ to $U^{(2)}$ from SAPT may make sense, adding it to $U^{(2)}$ from SAPT(DFT) leads to an overestimate of the interaction energy by about 4%. On the other hand, adding $U^{(3-\infty)}$(KS) to $U^{(2)}$ from SAPT(DFT) leads to an interaction energy *under*estimated by about 4%. Therefore it is possible that the $\delta^{\text{HF}}_{\text{int,resp}}$ correction is an overestimate of the higher-order terms.

The effects of these approximations on the total interaction energy for the water dimer are more clearly represented in Figure 2. The SAPT(DFT) interaction energy at second order, $U^{(2)}$[SAPT(DFT)], results in a potential that is clearly too shallow, with the repulsive wall and minimum both moved out toward larger *R*. Adding $\delta^{\text{HF}}_{\text{int,resp}}$ to $U^{(2)}$[SAPT(DFT)] results in a potential curve that is apparently too deep with both the repulsive wall and minimum moved inward. This potential curve is quite similar to the SAPT curve. The best agreement with the CCSD(T) potential is obtained with the $U^{(2)}$ [SAPT(DFT)]+$U^{(3-\infty)}$(KS) approximation, but this could be due to the slow convergence of the CCSD(T) interaction energy with respect to basis set.

Also reported in Table 1 are interaction energies for the strongly polar hydrogen fluoride dimer and the $H_2O\cdots H_3N$ dimer in a weakly polar geometry. The hydrogen fluoride dimer is probably the worst case for perturbation theory. For SAPT(DFT), $U^{(2)}$ constitutes only 83% of the reference CCSD(T) interaction energy and including $U^{(3-\infty)}$(KS) results in an improvement but still recovers only 91% of the reference. On the other hand, using the $\delta^{\text{HF}}_{\text{int,resp}}$ estimate of higher-order energies results in a near perfect agreement with CCSD(T). In contrast, for the $H_2O\cdots H_3N$ dimer, perturbation theory is rather rapidly convergent with $U^{(2)}$[SAPT(DFT)] and CCSD(T) differing by about 1% only. This good agreement is made slightly worse by the addition of the higher-order energy estimates as $U^{(2)}$[SAPT(DFT)]+ $\delta^{\text{HF}}_{\text{int,resp}}$ and $U^{(2)}$[SAPT(DFT)] + $U^{(3-\infty)}$(KS) differ from CCSD(T)

Table 1. Contribution of Third- and Higher-Order Corrections to the Interaction Energy for the Water, Hydrogen Fluoride, Carbon Dioxide, and Benzene Dimers and the $H_2O\cdots H_3N$ and $H_2\cdots CO$ Complexes[e]

method	energy component	$(H_2O)_2$	$(HF)_2$	$H_2O\cdots H_3N$	$(CO_2)_2$	$(C_6H_6)_2$	$H_2\cdots CO$
SAPT	$U^{(2)}$	−18.08	−15.59	−5.803	−7.05		−1.123
	$E^{(30)}_{\mathrm{ind,tot}}$	−0.80	−0.95	−0.017	−0.12		−0.018
	$\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$	−2.82	−3.14	−0.098	−0.23		−0.187
	$U^{(2)}+\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$	−20.90	−18.73	−5.901	−7.28		−1.310
SAPT(DFT)	$U^{(2)}$	−18.44	−15.38	−5.776	−5.72	−8.11[c]	−1.066
	$E^{(3)}_{\mathrm{ind,tot}}(\mathrm{KS})$	−1.20	−1.41	−0.027	−0.17	+0.29[d]	−0.023
	$E^{(4-\infty)}_{\mathrm{ind,d-class}}$	−0.04	−0.02	0.00	0.00	0.00	0.00
	$U^{(2)}+E^{(3)}_{\mathrm{ind,tot}}(\mathrm{KS})+E^{(4-\infty)}_{\mathrm{ind,d-class}}$	−19.67	−16.81	−5.803	−5.90	−7.82	−1.090
	$U^{(2)}+\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$	−21.26	−18.52	−5.874	−5.95		−1.253
CCSD(T)	$U^{\mathrm{CCSD(T)}}$	−20.45	−18.50	−5.694	−5.94[a]	−7.57[b]	−1.063

[a] Misquitta et al.[12] Dispersion optimized basis augmented with mid-bond functions. [b] Sinnokrot et al.[37] Estimate of the complete basis set CCSD(T) energy. [c] $E^{(2)}_{\mathrm{disp}}$ was calculated in a TZ/MC$^+$ basis set , the rest of the interaction energy components in the Sadlej/MC$^+$ basis set. [d] Computed using a Sadlej/MC$^+$ basis set. [e] The first three dimers are at their equilibrium geometries, the benzene dimer is in the parallel stacked geometry with a center-of-mass separation of 3.8 Å. $H_2\cdots CO$ is in the linear geometry with C toward H_2 and a center-of-mass separation of 7.8 au, and $H_2O\cdots H_3N$ is in geometry (a) of Figure 5 from ref 36. As stated in section III, the induction energy, $E^{(n)}_{\mathrm{ind,tot}}$, is defined as the *sum*: $E^{(n)}_{\mathrm{ind,pol}}+E^{(n)}_{\mathrm{ind,exch}}$. Unless otherwise specified, interaction energies were calculated using the aug-cc-PVTZ/MC$^+$ basis, and molecular properties needed for the damped classical model were obtained using the aug-cc-pVTZ/MC basis. All energies are reported in kJ mol^{-1}.

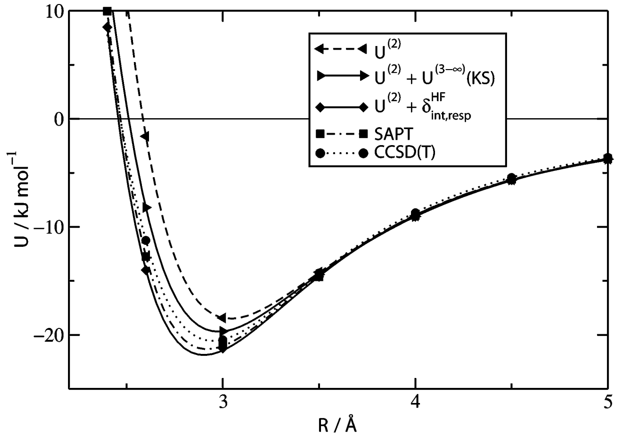


Figure 2. Total interaction energies of the water dimer obtained with SAPT(DFT) using different approximations to the higher-order energies. SAPT and CCSD(T) interaction energies are also shown. We have used the standard SAPT definition of the interaction energy that includes the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ correction (see, for example, eqs 4 and 5 in ref 12). All calculations were performed using the aug-cc-pVTZ basis in the MC$^+$ format. The dimer geometry is the same as used in Figure 1.

by 3% and 2%, respectively. Since these differences are small enough to be ascribed to basis incompleteness effects, we conclude that either estimate of the higher-order energies is suitable for this system.

We now turn to dimers for which the induction energy plays a relatively unimportant role. In Table 1 we report interaction energies for three such dimers: the carbon dioxide dimer, the benzene dimer, and the $H_2\cdots CO$ dimer. Results are far more encouraging for the nonpolar dimers. For the carbon dioxide dimer, all estimates of $U^{(3-\infty)}$ are very small but still constitute about 3% of the interaction energy. $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ and $U^{(3-\infty)}$(KS) are both about −0.2 kJ mol^{-1}, and SAPT(DFT) interaction energies obtained using either estimate of the higher-order energies results in an interaction energy almost identical with the CCSD(T) reference. For this particular system, $U^{(2)}$ from SAPT is too large in magnitude due to a severe overestimation of the dispersion energy made by the current implementation of SAPT.[12]

The benzene dimer is a rather unusual system as the higher-order induction effects are positive in the stacked geometry. Due to its size, no SAPT calculation was possible for this system. Indeed, accurate reference energies for this system are hard to obtain. The best reference energies currently available are those from Sinnokrot et al.,[37] obtained using R12-MP2 energies together with a CCSD(T) correction obtained using an aug-cc-pVDZ basis. Unfortunately, since the MP2 interaction energy is particularly poor for this system, being an overestimation by nearly 100%, the estimate of Sinnokrot et al. is probably in error by a few percent, and we should keep this in mind in the following comparisons. $U^{(2)}$ from SAPT(DFT) differs from the estimated CCSD(T) interaction energy differ by −7%. On adding the $U^{(3-\infty)}$(KS), this difference is reduced to only −3%. The higher-order contributions constitute about 4% of the total interaction energy of the dimer. Basis set incompleteness effects, estimated by Podeszwa et al.,[39] are of the same magnitude but of opposite sign. Therefore, the higher-order contributions to the interaction energy cannot be neglected in high-accuracy calculations, even for nonpolar systems, particularly as it is becoming usual to estimate the complete basis set limit of the interaction energies.

The $H_2\cdots CO$ dimer is the most weakly bound system included in Table 1. SAPT(DFT) converges very quickly to the CCSD(T) reference energy with $U^{(2)}$ [SAPT(DFT)] and CCSD(T) being essentially equal and $U^{(2)}$[SAPT(DFT)]+$U^{(3-\infty)}$(KS) differing from CCSD(T) by only 3%. In contrast, $U^{(2)}$-[SAPT(DFT)]+ $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ overshoots the CCSD(T) reference by 18%. The situation is worse for SAPT energies: $U^{(2)}$-[SAPT] is already too negative by about 6%, and the inclusion of the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ term makes the SAPT energy about 23% too negative. Clearly then, in agreement with Patkowski et al.,[20] the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ term is a very poor estimate of the higher-order energies for this dimer. This is the only system for which we have found this to be the case though Patkowski et al.[20] have found that $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ behaves likewise for the argon dimer.

From Table 1 and Figure 1 we see that for all of these dimers, the damped classical estimate of the induction effects above third order, $E^{(4-\infty)}_{\mathrm{ind,d-class}}$, is very small and could be neglected without a significant loss in accuracy. Thus our recommended approximation for the higher-order contributions to the interaction energy of a dimer is

$$U^{(3-\infty)} \approx U^{(3-\infty)}(\mathrm{KS}) \approx E^{(3)}_{\mathrm{ind,tot}}(\mathrm{KS}) \qquad (25)$$

IV.2. Many-Body Induction. In the condensed phase of water, the two-body interaction energies have been found to account for only about 85% of the total interaction energy per molecule.[23] The remaining 15% arises from many-body nonadditive effects, that is, that part of the interaction energy that cannot be represented by the sum of pairwise interactions. This nonadditivity is responsible for some of the important structural properties of water and, in particular, has a large role to play in hydrogen bonding. The nonadditive effects are even larger in small water clusters where they constitute between 17% and 30% of the total interaction energy.[22,24] These effects are expected to be equally important for polar molecules other than water and must be included in atom−atom potentials for organic molecules, which are commonly very polar and often form hydrogen-bonded networks.

The perturbative treatment of nonadditivity is a complex field of research, and while there is a version of SAPT that includes the three-body nonadditivity (see ref 7 for a review), the computational demands are so high as to preclude its applications to organic molecules. However, one of the major conclusions of accurate studies on water clusters[22−24] has been that the bulk of the nonadditivity for polar systems can be recovered using the relatively simple damped classical polarization model (see ref 3 for a description). While the exchange nonadditivity is not negligible for small clusters,[22,24] it is less important in relative terms for large clusters and the condensed phase, because additional coordination shells around any given molecule increase the dispersion, induction, and electrostatic energies but not the short-range contributions like the exchange nonadditivity.[23]

The damped classical polarizable model for a cluster of molecules is again given by eqs 21 and 22, but we now take into account all the molecules in the cluster. While the effect of the iterations in eq 22 is quite small for the two-body energy, iterations have a much larger role to play in larger clusters of polar molecules. For example, in clusters of water molecules optimized using the ASP-W4 potential,[22] iterations in eq 22 stabilize the dimer, trimer, tetramer, and pentamer by 2.5%, 6.0%, 10.4%, and 12.3% of the total interaction energy, respectively. These effects are clearly considerable. Furthermore, without iterations, many-body contributions to the interaction energy above the 3-body energies are absent (the $(3 + m)$-body terms arise at the mth iteration of eq 22). Work on organic crystals also indicates that iterations contribute strongly to the lattice energy.[40]

IV.3. Asymptotic Induction Energies. The asymptotic expansion of the two-body, second-order induction energy involves polarizability tensors and multipole moments of the unperturbed monomers (see ref 7 for a review). For large molecules, these molecular properties must be distributed, that is, expressed in terms of multiple sites—usually chosen to be the atomic centers—so as to improve the convergence of the multipole expansion. The distribution of the multipole moments has been the subject of many decades of research (see refs 3 and 41 for reviews). The Distributed Multipole Analysis (DMA) of Stone[42] is widely used, and a recent modification[43] overcomes a shortcoming of the earlier method that arose when diffuse functions were present in the basis. The problem of distributing the polarizabilities has proved harder, and only fairly recently have methods become available that are suitable for molecules of 20−30 atoms and modern basis sets, while being general enough to be applicable to frequency-dependent polarizabilities.[44,45]

The frequency-dependent polarizability, $\alpha_{lm,l'm'}(\omega)$, can be defined in terms of the FDDS as

$$\alpha_{lm,l'm'}(\omega) = \int\int \alpha(\mathbf{r}, \mathbf{r}'|\omega)\hat{Q}_{lm}(\mathbf{r})\hat{Q}_{l'm'}(\mathbf{r}')\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \qquad (26)$$

Real-space partitioning schemes have been based on ways of defining the multipole moment operators in the above expression so that they act on finite regions of space, defined, for example, by using integration grids[46] or by Bader's theory of atoms in molecules.[47] Both of these methods have shortcomings[45] that make them unsuitable for practical use. In contrast, the distribution scheme of Misquitta and Stone[45] focuses on a partitioning of the FDDS. The FDDS, as defined by eq 5, cannot be directly partitioned as the molecular orbitals that appear in this expression are generally delocalized. Rather, density-fitting[48,49] is used to simplify the form of the FDDS and then achieve the necessary site−site partitioning.

If the auxiliary basis set used to obtain the density-fitted FDDS in eq 10 is partitioned into contributions from individual sites, that is, $\{\chi\} = \{\chi^{(1)}, \chi^{(2)}, ...\}$, then the FDDS can be written as

$$\alpha(\mathbf{r},\mathbf{r}'|\omega) \approx \sum_{ab} \alpha^{a,b}(\mathbf{r}, \mathbf{r}'|\omega) \qquad (27)$$

where

$$\alpha^{a,b}(\mathbf{r}, \mathbf{r}'|\omega) = \sum_{p\in a,q\in b} \tilde{C}_{pq}(\omega)\chi_p(\mathbf{r})\chi_q(\mathbf{r}') \qquad (28)$$

Finally, using eqs 27 and 26 the distributed polarizability for sites (a, b) is defined as

$$\alpha_{lm,l'm'}{}^{a,b}(\omega) = \sum_{p\in a,q\in b} \tilde{C}_{pq}(\omega)N^p_{lm}N^q_{l'm'} \qquad (29)$$

where $N^p_{lm} = \int\hat{Q}_{lm}(\mathbf{r} - \mathbf{a})\chi_p(\mathbf{r})\mathrm{d}^3\mathbf{r}$, where $\boldsymbol{a}$ is a suitable reference origin for site a that will typically be taken to be the nucleus.

The standard density-fitting procedure[48,49] involves the minimization of the function

$$\Delta_{iv} = \int\int[\rho_{iv}(\mathbf{r}_1) - \tilde{\rho}_{iv}(\mathbf{r}_1)]\frac{1}{r_{12}}[\rho_{iv}(\mathbf{r}_2) - \tilde{\rho}_{iv}(\mathbf{r}_2)]\mathrm{d}^3\mathbf{r}_1\mathrm{d}^3\mathbf{r}_2 \qquad (30)$$

where the transition density $\rho_{iv} = \phi_i\phi_v$ is approximated by $\tilde{\rho}_{iv} = \sum_k D_{iv,k}\chi_k$. For the method of distribution based on eq 29 to work, it has been shown[45] that the function to be minimized must be replaced by one involving additional constraints

$$\Xi_{iv} = \Delta_{iv} - \eta \sum_{a,b\neq a} E_{iv}^{ab} + \lambda\left(\sum_a \sum_{k\in a} D_{iv,k} I_k\right)^2 \tag{31}$$

where η and λ are constants empirically determined to be about 0.0005 and 1000.0, respectively, $I_k = \int \chi_k(\mathbf{r})\mathrm{d}^3\mathbf{r}$, so that the last term imposes the orthogonality of the occupied and virtual orbitals (cf. eq 9), and E_{iv}^{ab} is the Coulomb interaction between the contributions of the basis functions of sites a and b to the transition density ρ_{iv} and is defined as

$$E_{iv}^{ab} = \int\int \frac{\tilde{\rho}_{iv}^{a}(\mathbf{r}_1)\tilde{\rho}_{iv}^{b}(\mathbf{r}_2)}{r_{12}} \mathrm{d}^3\mathbf{r}_1 \mathrm{d}^3\mathbf{r}_2 \tag{32}$$

The distributed polarizabilities obtained using this constrained density-fitting procedure contain nonlocal terms, that is, terms involving pairs of distinct sites. In contrast to other distribution methods, the nonlocal terms describing flow of charge from site to site are very small (around 10^{-2} and 10^{-3} in magnitude) for all systems, irrespective of the type of bonding involved. Nevertheless, nonlocal terms are best avoided as they complicate the description unnecessarily. In ref 45, the Le Sueur and Stone localization method[50] was used to transform the nonlocal terms into local polarizabilities and remove the charge-flow terms altogether. The localization by this procedure causes a deterioration of the convergence properties of the model, because multipole expansions are used to move the polarizabilities around. In principle, this can increase the radius of divergence of the description to be equal to the size of the molecule, thereby causing significant losses in accuracy for large molecules. Thus, while good results have been obtained for molecules like formamide and urea, there is already an appreciable loss in accuracy for *N*-methyl propanamide.[45]

In ref 45 it was suggested that the Williams and Stone method[44] of obtaining local polarizabilities using a fit to the point-to-point polarizabilities could be used to refine the polarizability model obtained from the constrained density-fitting procedure. This can be done as follows. The point-to-point polarizability α_{PQ} describes the response of the electrostatic potential at a point $\boldsymbol{Q}$ to the frequency-dependent potential produced by a unit oscillating point charge at point $\boldsymbol{P}$ and is given by the expression

$$\alpha_{PQ}(\omega) = \int\int \alpha(\mathbf{r},\mathbf{r}'|\omega)\hat{O}^P(\mathbf{r})\hat{O}^Q(\mathbf{r}')\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \tag{33}$$

where $\hat{O}^P(\mathbf{r}) = -q_P/(4\pi\epsilon_0|\boldsymbol{P}-\mathbf{r}|)$ and $\hat{O}^Q(\mathbf{r}) = -q_Q/(4\pi\epsilon_0|\boldsymbol{Q} - \mathbf{r}|)$. These polarizabilities are evaluated on a random grid of points typically between the vdW × 2 and vdW × 4 surfaces. These responses can be evaluated very efficiently using the density-fitted form of eq 33 given in ref 45. For a grid of N points, where N is typically a few thousand, there are $^1/_2N(N + 1)$ responses, which can be obtained in a single calculation. In the original Williams and Stone method, a model is postulated comprising polarizabilities α_{tu}^{ab}, where t = 00, 10, 10*c*, 10*s*, ..., . In terms of this model the responses are given by

$$\tilde{\alpha}_{PQ} = \sum_{ab}\sum_{tu} T_{0t}^{Pa}\,\alpha_{tu}^{ab}\,T_{u0}^{bQ} \tag{34}$$

where T_{0t}^{Pa} is the interaction tensor[3] which describes the interaction between a point charge at P and a multipole Q_t at point a. The model polarizabilities α_{tu}^{ab} are then obtained by minimizing the squared difference

$$S = \sum_{PQ} (\tilde{\alpha}_{PQ} - \alpha_{PQ})^2 \tag{35}$$

This procedure is very accurate and leads to a compact description of the polarizability, but the resulting polarizabilities are not always positive definite. This tends to happen for 'buried' atoms, i.e., atoms hidden under the van der Waals spheres of neighboring atoms, and could, in principle, lead to positive induction energies, which is physically impossible.

These unphysical terms can be avoided by using the models obtained using the localized polarizabilities from the constrained density-fitting method as 'anchor' values and minimizing

$$S = \sum_{PQ} (\tilde{\alpha}_{PQ} - \alpha_{PQ})^2 + \sum_{kk'} g_{kk'}(p_k - p_k^0)(p_{k'} - p_{k'}^0) \tag{36}$$

where the p_k are the parameters in the model, i.e., the α_{tu}^{ab} defined above, p_k^0 are the 'anchor' values, and $g_{kk'}$ are elements of a positive definite matrix that could be taken to be diagonal.

This combination of the Williams−Stone and Misquitta−Stone procedures will be called the Williams−Stone−Misquitta (WSM) procedure. Initial results using the WSM procedure have been very encouraging and have been presented in a recent review article.[41] More extensive results will be presented in part 2.

V. Summary

In this first part of our investigation, we have laid down the theoretical framework for the accurate calculation of the induction energy of clusters of organic molecules.

We have broken away from convention by identifying the induction energy with the *sum* of the second-order induction energy as defined through the polarization expansion,[6,7] termed $E_{\text{ind,pol}}$, and its exchange counterpart, the exchange-induction energy $E_{\text{ind,exch}}$. Thus, we define the nth order induction energy as $E_{\text{ind,tot}}^{(n)} = E_{\text{ind,pol}}^{(n)} + E_{\text{ind,exch}}^{(n)}$. This definition was motivated by both theoretical and numerical considerations.

The two-body induction energy at second order in the interaction operator, $E_{\text{ind,pol}}^{(2)}$, is the most important contribution of the induction energy to the interaction energy of a cluster of molecules. We have presented a density-fitted form of the SAPT(DFT) expression for $E_{\text{ind,pol}}^{(2)}$ that is both accurate and computationally efficient. This is a natural extension of the density-fitting technique that one of us has used for the dispersion energy[10,12] and that has already been proposed—in a different form—by Hesselmann et al.[16] As well as a reduction in the computational cost from $O(n_o^2 n_v^2)$ to $O(m^2)$ where n_o and n_v are the number of occupied and virtual orbitals, respectively, and m is the number of auxiliary basis functions, we gain by avoiding the computationally expensive 4-index Coulomb integrals needed to evaluate the original SAPT(DFT) expression. The new formulation has been implemented in the CamCASP program[51] which was used for all calculations of $E^{(2)}_{\mathrm{ind,pol}}$ reported in this paper.

The two-body interaction energy has major contributions from energies of third and higher order in the interaction operator. These higher-order energies are predominantly induction in nature and can contribute as much as 17% of the two-body interaction energy for hydrogen-bonded complexes. They have usually been estimated using the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ correction[7,17,18] which is cumbersome to calculate as it involves a supermolecular Hartree−Fock calculation of the induction energy and a low-order SAPT calculation. We have proposed that these energies be approximated using the third-order induction energy calculated using SAPT(KS), i.e., $E^{(3)}_{\mathrm{ind,tot}}$(KS). Since SAPT(KS) is the first step in a SAPT(DFT) calculation of the interaction energy, this entails little additional effort.

$E^{(3)}_{\mathrm{ind,tot}}$(KS) has been shown to approximate the higher-order energies rather well for non-hydrogen-bonded dimers, where we get almost perfect agreement with the reference interaction energies, but it underestimates them for hydrogen-bonded dimers. For example, the interaction energy calculated using this approximation is underestimated by around 4% for the water dimer and 9% for the hydrogen fluoride dimer, both at their global minimum geometries. The hydrogen fluoride dimer is probably the worst case for methods based on perturbation theory as higher-order energies are estimated to constitute about 17% of the interaction energy at the global minimum geometry. For the non-hydrogen-bonded dimers, the higher-order energies are smaller but still constitute about 4% of the interaction energy. Patkowski et al.[20] were led to similar conclusions in an investigation of the SAPT interaction energy.

With the exception of the very weakly bound $H_2\cdots CO$ dimer, for the polar and nonpolar dimers studied here, we found that the $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ correction provides a reasonably accurate estimate of higher-order energies. This conclusion complements that of Patkowski et al.[20] It is quite possible that $\delta^{\mathrm{HF}}_{\mathrm{int,resp}}$ is indeed a poor estimate of the higher-order energies for very weakly bound dimers but might be more reasonable for the more strongly bound dimers, whether polar of not. More data from a larger variety of systems will be needed to test this conjecture.

Yet another way of estimating the higher-order energies is through the damped classical polarizable model. The classical estimate for the higher-order energy is obtained by iterating the fields and induced multipoles self-consistently to convergence. It is necessary to use a distributed-polarizability description for all but the smallest molecules, and the Williams−Stone−Misquitta procedure[44,45] provides an efficient and accurate route to such descriptions. (See results provided in part 2.) We have found that the damped classical polarizable model severely underestimates the higher-order energies for the two-body interaction. Recall that it is incomplete because it assumes linear response of each molecule to external fields and neglects orbital overlap effects. For example, it recovers less than 10% of the interaction energy contribution from third and higher orders for the water dimer at its global minimum geometry. Therefore, using the damped classical model to estimate these energies would result in an error of about 9%, or about 1.8 kJ mol^{-1}, in the total interaction energy of the water dimer.

However, the effect of the iterations can be quite large in clusters of polar molecules. For example, the additional stabilization is about 12% of the total interaction energy for the water pentamer.[22] Iterations have also been shown to make major contributions to the lattice energy of organic crystals.[40] Therefore, we do recommend that the iterated form of the damped classical polarizable model be used in calculations involving polar clusters.

We should perhaps emphasize that the higher-order energies are *larger* in magnitude than the basis set incompleteness errors in the SAPT(DFT) calculations. Therefore these energies cannot be ignored in accurate calculations, especially those attempting to estimate the complete-basis-set energy.

Our recommended expression for the total interaction energy of a dimer calculated using SAPT(DFT)[12] is

$$U \approx E^{(1)}_{\mathrm{elst}}(\mathrm{KS}) + E^{(1)}_{\mathrm{exch}}(\mathrm{KS}) + E^{(2)}_{\mathrm{ind,tot}} + E^{(2)}_{\mathrm{disp}} + E^{(2)}_{\mathrm{disp,exch}} + U^{(3-\infty)} \quad (37)$$

where $E^{(1)}_{\mathrm{elst}}$(KS) and $E^{(1)}_{\mathrm{exch}}$(KS) are the first-order electrostatic and exchange energies, respectively, $E^{(2)}_{\mathrm{disp}}$ and $E^{(2)}_{\mathrm{disp,exch}}$ are the second-order dispersion and exchange-dispersion energies, respectively, $E^{(2)}_{\mathrm{ind,tot}}$ is given by eq 19, and $U^{(3-\infty)}$ is approximated as in eq 25. This approximation does not include higher-order dispersion terms or nonlinear induction effects. In Part 2 we will explore further approximations and also investigate the numerical issues associated with calculations of the induction energy of dimers and clusters of organic molecules.

Acknowledgment. A.J.M. would like to thank Girton College, Cambridge for a research fellowship. This research was supported by EPSRC grant EP/C539109/1. We are grateful to a referee for very insightful and helpful comments.

References

(1) Misquitta, A. J.; Stone, A. J.; Price, S. L. Accurate induction energies for small organic molecules: 2. Models and numerical details. *J. Chem. Theory Comput.* **2007**, *3*, 19−32.

(2) Bukowski, R.; Szalewicz, K.; Groenenboom, G.; van der Avoird, A. Interaction potential for water dimer from symmetry-adapted perturbation theory based on density functional description of monomers. *J. Chem. Phys.* **2006**, *125*, 044301.

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

(4) Magnasco, V.; McWeeny, R. Weak interaction between molecules and their physical interpretations. In *Theoretical Models of Chemical Bonding*; Maksić, Z. B., Ed.; Springer: New York, 1991; Vol. 4, pp 133−169.

(5) McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd ed.; Academic Press: New York, 1992.

(6) 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.

(7) Jeziorski, B.; Szalewicz, K. Symmetry-adapted perturbation theory. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley: 2002; Vol. 8, pp 37−83.

(8) Szalewicz, K.; Patkowski, K.; Jeziorski, B. Intermolecular interactions via perturbation theory: From diatoms to biomolecules. In *Intermolecular Forces and Clusters II*; Wales, D. J., Ed.; Springer-Verlag: Berlin, Heidelberg, 2005; Vol. 116 of *Structure and Bonding*, pp 43−117.

(9) Misquitta, A. J.; Szalewicz, K. Intermolecular forces from asymptotically corrected density functional description of monomers. *Chem. Phys. Lett.* **2002**, *357*, 301−306.

(10) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **2003**, *91*, 33201.

(11) Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.* **2005**, *122*, 214109.

(12) 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 theory. *J. Chem. Phys.* **2005**, *123*, 214103.

(13) Hesselmann, A.; Jansen, G. First-order intermolecular interaction energies from Kohn−Sham orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464−470.

(14) 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.

(15) Hesselmann, A.; Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **2003**, *367*, 778−784.

(16) 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) Jeziorska, M.; Jeziorski, B.; Cizek, J. Direct calculation of the Hartree−Fock interaction energy via exchange perturbation expansion—the He···He interaction. *Int. J. Quantum Chem.* **1987**, *32*, 149−164.

(18) Moszynski, R.; Heijmen, T. G. A.; Jeziorski, B. Symmetry-adapted perturbation theory for the calculation of Hartree−Fock interaction energies. *Mol. Phys.* **1996**, *88*, 741−758.

(19) Mas, E. M.; Bukowski, R.; Szalewicz, K. Ab initio three-body interactions for water. I. Potential and structure of water trimer. *J. Chem. Phys.* **2003**, *118*, 4386−4403.

(20) Patkowski, K.; Szalewicz, K.; Jeziorski, B. Third-order interactions in symmetry-adapted perturbation theory. *J. Chem. Phys.* **2006**, *125*, 154107.

(21) Patkowski, K.; Jeziorski, B.; Szalewicz, K. Symmetry-adapted perturbation theory with regularized coulomb potential. *J. Mol. Struct. (THEOCHEM)* **2001**, *547*, 293−307.

(22) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. Contribution of many-body terms to the energy for small water clusters: A comparison of ab initio calculations and accurate model potentials. *J. Phys. Chem. A* **1997**, *101*, 9163−9168.

(23) Mas, E. M.; Bukowski, R.; Szalewicz, K. Ab initio three-body interactions for water. II. Effects on structure and energetics of liquid. *J. Chem. Phys.* **2003**, *118*, 4404−4413.

(24) Milet, A.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. Hydrogen bonding in water clusters: Pair and many-body interactions from symmetry-adapted perturbation theory. *J. Phys. Chem. A* **1999**, *103*, 6811−6819.

(25) Casida, M. E. Time-dependent density-functional response theory for molecules. In *Recent Advances in Density-Functional Theory*; Chong, D. P., Ed.; World Scientific: 1995; p 155.

(26) Colwell, S. M.; Handy, N. C.; Lee, A. M. Determination of frequency-dependent polarizabilities using current density-functional theory. *Phys. Rev. A* **1996**, *53*, 1316−1322.

(27) Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. Excitation energies from time-dependent density-functional theory. *Phys. Rev. Lett.* **1996**, *76*, 1212−1215.

(28) Hesselmann, A.; Jansen, G. First-order intermolecular interaction energies from Kohn−Sham orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464−470.

(29) Tozer, D. J.; Handy, N. C. Improving virtual Kohn−Sham orbitals and eigenvalues: Application to excitation energies and static polarizabilities. *J. Chem. Phys.* **1998**, *109*, 10180−10189.

(30) Tozer, D. J. The asymptotic exchange potential in Kohn−Sham theory. *J. Chem. Phys.* **2000**, *112*, 3507−3515.

(31) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(32) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Density-fitting method in symmetry-adapted perturbation theory based on Kohn−Sham description of monomers. *J. Chem. Theory Comput.* **2006**, *2*, 400−412.

(33) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew−Burke−Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029−5036.

(34) Misquitta, A. J.; Stone, A. J. Regularized SAPT(DFT) **2007**, manuscript in preparation.

(35) Mas, E. M.; Szalewicz, K.; Bukowski, R.; Jeziorski, B. Pair potential for water from symmetry-adapted perturbation theory. *J. Chem. Phys.* **1997**, *107*, 4207−4218.

(36) Langlet, J.; Caillet, J.; Bergès, J.; Reinhardt, P. Comparison of two ways to decompose intermolecular interactions for hydrogen-bonded dimer systems. *J. Chem. Phys.* **2003**, *118*, 6157−6166.

(37) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. Estimates of the ab initio limit for $\pi-\pi$ interactions: The benzene dimer. *J. Am. Chem. Soc.* **2002**, *124*, 10887−10893.

(38) Moszyński, R.; Cybulski, S. M.; Chałasiński, G. Many-body theory of intermolecular induction interactions. *J. Chem. Phys.* **1994**, *100*, 4998−5010.

(39) Podeszwa, P.; Bukowski, R.; Szalewicz, K. Potential energy surface for the benzene dimer and perturbational analysis of $\pi-\pi$ interactions. *J. Phys. Chem. A* **2006**, *110*, 10345−10354.

(40) Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. Is the induction energy important for modelling organic crystals? *J. Chem. Theory Comput.* **2007**, manuscript in preparation.

(41) Stone, A. J.; Misquitta, A. J. Atom−atom potentials from *ab initio* calculations. *Int. Rev. Phys. Chem.* **2007**, *26*, 193−222.

(42) Stone, A. J.; Alderton, M. Distributed multipole analysis—methods and applications. *Mol. Phys.* **1985**, *56*, 1047−1064.

(43) Stone, A. J. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.

(44) Williams, G. J.; Stone, A. J. Distributed dispersion: a new approach. *J. Chem. Phys.* **2003**, *119*, 4620−4628.

(45) Misquitta, A. J.; Stone, A. J. Distributed polarizabilities obtained using a constrained density-fitting algorithm. *J. Chem. Phys.* **2006**, *124*, 024111.

(46) Le Sueur, C. R.; Stone, A. J. Practical schemes for distributed polarizabilities. *Mol. Phys.* **1993**, *78*, 1267−1291.

(47) Angyan, J. G.; Jansen, G.; Loos, M.; Hattig, C.; Hess, B. A. Distributed polarizabilities using the topological theory of atoms in molecules. *Chem. Phys. Lett.* **1994**, *219*, 267−273.

(48) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. On first-row diatomic molecules and local density models. *J. Chem. Phys.* **1979**, *71*, 4993−4999.

(49) Dunlap, B. I. Robust and variational fitting. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2113−2116.

(50) Le Sueur, C. R.; Stone, A. J. Localization methods for distributed polarizabilities. *Mol. Phys.* **1994**, *83*, 293−308.

(51) Misquitta, A. J.; Stone, A. J. *CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form*; University of Cambridge: 2006. Inquiries to A. J. Misquitta, am592@cam.ac.uk.

CT700104T

JCTC Journal of Chemical Theory and Computation

Accurate Induction Energies for Small Organic Molecules. 2. Development and Testing of Distributed Polarizability Models against SAPT(DFT) Energies

Alston J. Misquitta,[†,‡] Anthony J. Stone,*[,†] and Sarah L. Price[‡]

University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, U.K., and University College London, 20 Gordon Street, London WC1H 0AJ, U.K.

Received May 3, 2007

Abstract: In part 1 of this two-part investigation we set out the theoretical basis for constructing accurate models of the induction energy of clusters of moderately sized organic molecules. In this paper we use these techniques to develop a variety of accurate distributed polarizability models for a set of representative molecules that include formamide, *N*-methyl propanamide, benzene, and 3-azabicyclo[3.3.1]nonane-2,4-dione. We have also explored damping, penetration, and basis set effects. In particular, we have provided a way to treat the damping of the induction expansion. Different approximations to the induction energy are evaluated against accurate SAPT(DFT) energies, and we demonstrate the accuracy of our induction models on the formamide−water dimer.

I. Introduction

In this paper, which is the second part in a two-part investigation[1] of the induction energy, we report methods for developing models of the induction energy that are suitable for applications involving organic molecules. In the first part of our study, which we will refer to as part 1, we set out the theoretical framework for the calculation of the nonexpanded and expanded induction energies in a way that is suitable for clusters of organic molecules. We now use that framework to develop the distributed polarizability models needed to model the induction energy in the condensed phase and demonstrate how the refinement procedure described in part 1 can be used to obtain polarizability models that combine accuracy and computational simplicity, making them ideal for real-world applications, in particular those in the field of organic crystal structure prediction.

The dominant part of the induction contribution to the total interaction energy of a cluster of molecules arises from the two-body interactions. In part 1 we argued that the SAPT(DFT) two-body interaction energy should be defined as

$$U = E^{(1)}_{\rm elst}({\rm KS}) + E^{(1)}_{\rm exch}({\rm KS}) + E^{(2)}_{\rm ind,tot} + E^{(2)}_{\rm disp} + E^{(2)}_{\rm exch-disp} + U^{(3-\infty)} \quad (1)$$

where $E^{(1)}_{\rm elst}$ (KS) and $E^{(1)}_{\rm exch}$ (KS) are the first-order electrostatic and exchange energies, respectively, $E^{(2)}_{\rm disp}$ and $E^{(2)}_{\rm exch-disp}$ are the second-order dispersion and exchange-dispersion energies, respectively, and $E^{(2)}_{\rm ind,tot}$ and $U^{(3-\infty)}$ were defined in part 1. The induction contribution to the two-body interaction energy arises at second and higher orders in the interaction operator. The most computationally efficient and accurate method for the calculation of the two-body induction energy at second order is based on the recently developed symmetry-adapted perturbation theory based on density functional theory, called SAPT(DFT)[2−4] or DFT-SAPT.[5] Higher-order induction energies form the bulk of the higher-order effects for polar molecules and can be very important for hydrogen-bonded dimers, where they can contribute as much as 17% of the two-body energy. In part 1 we described how induction contributions to the two-body energy that arise from terms of third and higher order in the interaction operator can be approximated within SAPT(KS).[3,6] Since a SAPT(DFT) calculation of the interaction energy uses energies computed within SAPT(KS), we are now able to calculate accurate nonexpanded induction energies within one theoretical framework. Additionally, due to the modest

* Corresponding author phone: +44 1223 336375; fax: +44 1223 336362; e-mail: ajs1@cam.ac.uk.

[†] University Chemical Laboratory.

[‡] University College London.

10.1021/ct700105f CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/10/2007

computational resources needed for such a calculation, we are able to apply these methods to dimers of small organic molecules.

What remains then is for us to model the induction energy in a way that is suitable for calculations on clusters of polar molecules. In such clusters, many-body induction energies can be almost as important as the two-body induction energies. Both can be described using the damped classical polarizable model,[7] and in fact higher-order contributions to the many-body energy can also be obtained from this model. The latter, which can form as much as 12% of the total energy of the cluster, are obtained from the classical model by iterating the fields and the responses to the fields to self-consistency (see section IV.2 in part 1).

We need accurate molecular polarizabilities for the classical polarizable model. For all but the smallest molecules, they are needed in a distributed form. Recently two of us have developed a method of obtaining distributed polarizabilities based on a partitioning of the molecular transition densities using a constrained density-fitting technique that is both accurate and applicable to large molecules.[8] In ref 8 we observed that the accuracy of the model deteriorates if it is simplified to include only local polarizability terms, that is to omit terms involving pairs of sites. This deterioration can be overcome by refining the polarizability model by the method of Williams and Stone.[9] This two-step procedure has an advantage over the Williams and Stone method alone, which can lead to unphysical nonpositive-definite terms. In the combined procedure, such terms are completely removed from the low-ranking polarizabilities and greatly reduced in the higher-ranking terms. In a recent review[7] some preliminary results of this Williams−Stone−Misquitta (WSM) procedure were presented, and in part 1 we gave its theoretical basis in some detail. Here we describe the numerical details of the procedure and present extensions of the method that can be used to calculate local polarizability models up to rank 2. Rank 1, that is, dipole−dipole, polarizabilities may be sufficient for calculations of moderate accuracy, but the higher-rank dipole−quadrupole and quadrupole−quadrupole terms are needed to achieve higher accuracy. On the other hand, for applications that are so computationally demanding as to require a simpler model, the WSM procedure can provide the best accuracy subject to such constraints.

At short intermolecular separations the classical polarizable model will be in error, and the difference between the nonexpanded induction energies calculated using SAPT(DFT) and the energies from the classical model will need to be accounted for. Additionally, at short separations, the classical polarizable model, which is based on a multipole expansion, can result in divergent energies. This problem arises quite often in the condensed phase. We have analyzed the problems associated with short intermolecular separations in some detail and have tried to provide solutions to many of them.

This paper is organized as follows: In section II we describe a powerful graphical technique for displaying the model induction energies. In section III we discuss the features of basis sets that are needed to obtain accurate molecular polarizabilities and also SAPT(DFT) induction energies.

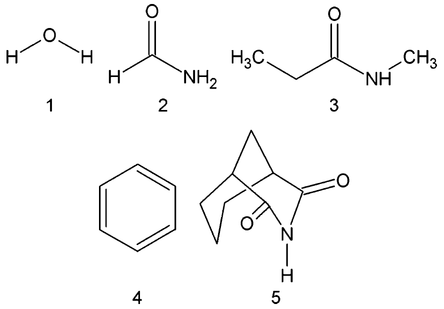


Figure 1. Molecules used to test distributed polarizability models: (1) water, (2) formamide, (3) *N*-methyl propanamide (N-MPA), (4) benzene, and (5) 3-azabicyclo[3.3.1]nonane-2,4-dione(BOQQUT).

The methods that we describe here for distributed polarizabilities can be used to obtain models of varying complexity and accuracy. In section IV we assess these models for the molecules shown in Figure 1, which have been chosen to provide a range of sizes and different types of charge distribution. Benzene tests the modeling of the polarizability of conjugated systems, in contrast to the saturated hydrocarbon functional groups. The other molecules give a range of hydrogen bond donor and acceptor strengths: the imide has a plastic phase[10] indicating that the hydrogen-bonding in its ordered polymorphs is readily disrupted.

In section V we discuss effects of penetration, truncation, and damping and propose a way to determine the damping coefficient based on molecular ionization energies. In section VI we assess approximations for calculating the induction energies. Finally, in section VII we conclude with a summary of the main results of this paper.

I.1. A Note on Notation. The notation we have used for the induction energies defined within SAPT(DFT) is somewhat nonstandard. We have described it in some detail in section III of part 1, but the key ideas are summarized here for convenience.

At order *n*, there are two components to the induction: the induction as defined through the polarization expansion,[11,12] termed $E^{(n)}_{\rm ind,pol}$, and the exchange component of the induction, termed $E^{(n)}_{\rm ind,exch}$. In part 1 we defined the *n*th order induction energy as the sum of these contributions, i.e., $E^{(n)}_{\rm ind,tot} = E^{(n)}_{\rm ind,pol} + E^{(n)}_{\rm ind,exch}$. The reasons for this definition, rather than the more conventional identification of $E^{(n)}_{\rm ind,pol}$ as the *n*th order induction, have been outlined in section III in part 1. In brief, our choice has been made because Coulomb singularities in the interaction operator mean that neither $E^{(n)}_{\rm ind,pol}$ nor $E^{(n)}_{\rm ind,exch}$ are meaningful on their own,[13,14] and the observation that the expanded induction energy, termed $E^{(n)}_{\rm ind,d-class}$, agrees best with $E^{(n)}_{\rm ind,tot}$ as defined here. Numerical evidence will be provided below.

II. Displaying the Energies

A powerful way to understand the various models that will be presented below is by mapping energies onto a suitable surface around the molecule in question. Such a mapping

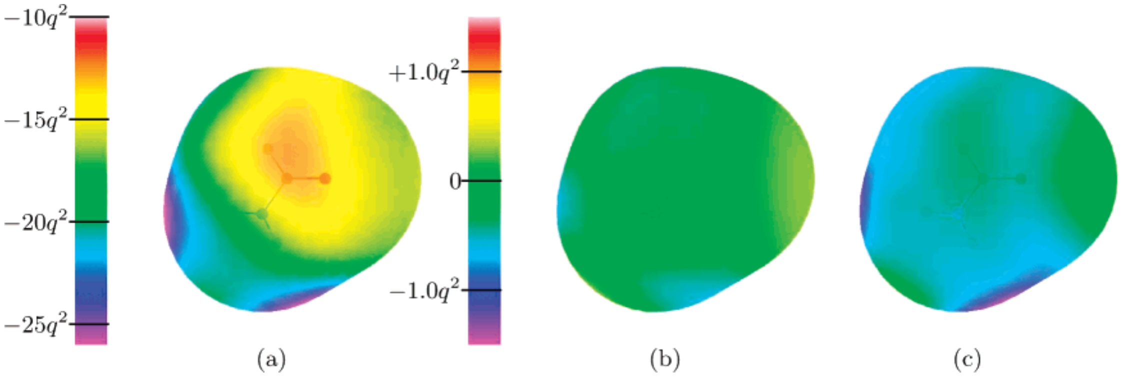


Figure 2. Induction energy map and difference maps (kJ mol^{-1}) arising from a charge q atomic units on the vdW × 2 surface of formamide obtained using a rank 4 nonlocal polarizability description. In (a) is displayed the induction energy map obtained using distributed polarizabilities calculated with the Sadlej basis set. Also shown are maps comparing the induction energies obtained using distributed polarizabilities calculated with the (b) aug-cc-pVTZ and (c) aug-cc-pVQZ bases with the Sadlej-basis energies.

can be most easily done if the energy probe has spherical symmetry, and a convenient probe for the induction energy is a point charge. The maps in this paper show the induction energies that result from a charge qe on the chosen surface. The SAPT(DFT) expression for the induction energy of a molecule in the field of a point charge is given by eq 17 in part 1. The induction energy of a molecule in the field of a point charge depends quadratically on the magnitude of the charge, and an appropriate value of q needs to be used in interpreting the energy scales in these maps. Setting $q = 1$ gives the response to a unit charge, but this is larger than typical local charges in a molecule, which are not expected to exceed 0.5e.

The surface around the molecule is constructed as follows. If the required distance from atom a is R_a^0 (e.g., twice the van der Waals radius for the vdW × 2 surface), then the surface is defined by $R_a - R_a^0 = 0$, or equivalently by $\exp[-\xi(R_a - R_a^0)] = 1$, where ξ is an arbitrary constant. We define the surface for the whole molecule by $\sum_a \exp[-\xi(R_a - R_a^0)] = 1$. The effect is as if we shrank an elastic membrane onto the union of vdW × 2 atomic surfaces, more or less tightly depending on the value of ξ; the intersections between the vdW surfaces of neighboring atoms are smoothed out. A value of $\xi = 2$ has been used for the maps shown here. The SAPT(DFT) maps were generated using the CamCASP program,[15] and the maps using distributed polarizabilities were generated with the ORIENT program, version 4.6.[16]

The van der Waals radii prescribed by Bondi[17] have been used for all but the hydrogen atoms that can form hydrogen bonds, which have their radii set to zero in order to reflect better the small interatomic distances associated with hydrogen bonds. The surface is then approximately the surface of contact for neighboring non-hydrogen atoms.

III. Choice of Basis

We will discuss the basis set requirements for obtaining accurate molecular properties and accurate SAPT(DFT) energies separately, because the requirements for large and small intermolecular separations are generally quite different. At short range orbital overlap effects become important and we generally need to supplement the basis sets used for SAPT(DFT) calculations with additional functions, while at long range the multipole expansion can be used and all we need are basis-saturated molecular properties.

III.1. Basis Set Requirements for Molecular Polarizabilities and Multipole Moments. The large size of even the smaller organic molecules makes the use of basis sets higher than triple-ζ quality hard to use on a routine basis. In fact, sufficiently accurate molecular polarizabilities are obtained with the Sadlej basis set[18,19] *if* used within the linear response Kohn−Sham DFT framework described in section IV of part 1. As an example, consider the formamide molecule. The induction map using distributed rank 4 nonlocal polarizabilities obtained with the Sadlej basis is displayed in Figure 2, together with the corresponding results for the aug-cc-pVTZ and aug-cc-pVQZ basis sets, displayed as *difference* maps against the Sadlej basis results. As would be expected, the differences are very small for the aug-cc-pVTZ basis, indicating that this basis is roughly equivalent to the Sadlej, at least for calculations of the induction energy. The differences are somewhat larger for the aug-cc-pVQZ basis, but, even in this case, the largest difference is around 1.5q^2 kJ mol^{-1} which is an order of magnitude less than the actual energies. For a more realistic charge of 0.5 units, the maximum difference would be about 0.4 kJ mol^{-1}.

We conclude that the Sadlej basis sets provide a very good compromise between size and accuracy for our purposes. They have been optimized for molecular properties but are about half the size of the equivalent aug-cc-pVTZ Dunning basis sets, thus significantly raising the limit on the size of molecules that can be used in such calculations.

As explained in section IV of part 1, we use density-fitting techniques in our calculations of the nonexpanded induction energy and distributed polarizabilities. As yet, there is no auxiliary density-fitting basis optimized for the Sadlej basis, so in view of the similarities between the Sadlej and aug-cc-pVTZ bases we have used the aug-cc-pVTZ auxiliary basis instead. We have confirmed that this is a good choice by carrying out extensive tests of molecular properties and interaction energies.

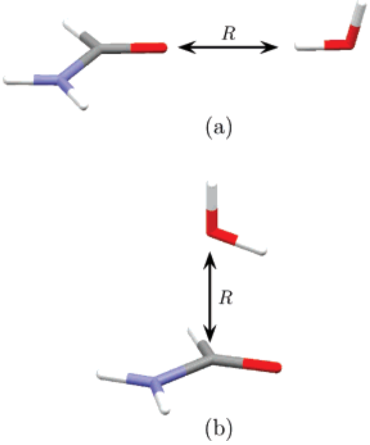


Figure 3. Formamide water dimer geometries used in this paper. Geometry (a) emphasizes the hydrogen oxygen contacts and geometry (b) emphasizes the contacts between the heavy atoms. The radial minima in the interaction energy for geometries (a) and (b) are at approximately 4 and 5.5 bohr, respectively.

III.2. Basis Set Requirements for the SAPT(DFT) Induction Energies. Monomer basis sets of triple-ζ quality are generally suitable for accurate SAPT(DFT) induction energies if the intermolecular separation is large and overlap effects are negligible. Basis sets with functions located only on the atomic sites of the monomer are said to be of the 'monomer centered' or MC type. For intermediate and short molecular separations where overlap effects are significant, basis-converged SAPT(DFT) energies are obtained only with basis sets supplemented with functions located on the nuclei of the interacting partners.[20] These are the so-called 'far-bond' functions which typically comprise just the *s* and *p* symmetry functions of the interacting monomer. To saturate the dispersion energy, a further small set of functions is needed in the region between the interacting molecules—the so-called 'mid-bond' functions. The resulting basis set is said to be of the MC$^+$ type, where the '+' sign indicates the presence of the additional basis functions. Although the mid-bond functions have a negligible effect on the induction energies, we will include them in all SAPT(DFT) calculations, for consistency with later work.

The effect of the far-bond functions on the polarization expression for the induction and exchange-induction energies *individually* is rather dramatic.[20] At the minimum-energy intermolecular separation, these energies can increase in magnitude by 2 orders of magnitude upon inclusion of the far-bond functions; but $E^{(n)}_{\text{ind,pol}}$ is significantly quenched by $E^{(n)}_{\text{ind,exch}}$, at every order n.[21] Consequently, as discussed in sections III and IV.1 of part 1, it is more useful to consider $E^{(n)}_{\text{ind,tot}}$, the sum of these two energies at each order.

In Figure 4 we display $E^{(2)}_{\text{ind,tot}}$ and $E^{(3)}_{\text{ind,tot}}$ calculated using the Sadlej basis in MC and MC$^+$ basis types for the formamide water dimer. The MC$^+$ energies are uniformly larger in magnitude (more negative) than the MC results. Near the minimum-energy separations, the difference between the MC$^+$ and MC results for geometry (a) is about 1 kJ mol^{-1} for $E^{(2)}_{\text{ind,tot}}$ and 0.7 kJ mol^{-1} for $E^{(3)}_{\text{ind,tot}}$. For geometry (b) the corresponding differences are about 0.5 and 0.2 kJ mol^{-1}. These changes constitute around 6% of the total interaction energies near the minimum-energy separations for the two configurations. These are not small effects and cannot be ignored in accurate studies. Perhaps more importantly from the point of view of geometry optimizations, the MC$^+$ basis sets result in deeper wells and smaller radial separations, particularly for the hydrogen-bonding geometries.

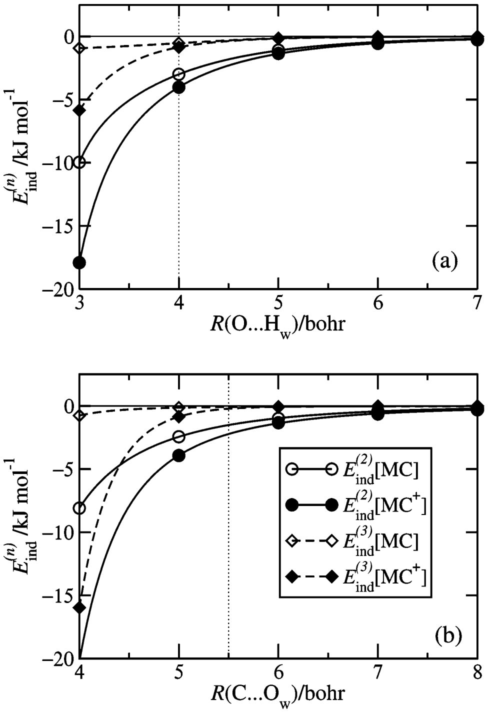


Figure 4. The effect of far-bond functions on the second- and third-order induction energies, respectively, for the formamide−water dimer. See Figure 3 for a description of the dimer geometries.

The above picture remains the same if the aug-cc-pVTZ basis is used in place of the Sadlej basis. The differences in $E^{(2)}_{\text{ind,tot}}$ and $E^{(3)}_{\text{ind,tot}}$ calculated using these two bases are already small (though not negligible) for the MC basis type, and with the MC$^+$ type, the aug-cc-pVTZ and Sadlej bases yield essentially the same energies. We should emphasize here that this does not mean that the *total* interaction energy can be calculated using the Sadlej/MC$^+$ basis. This is because the basis incompleteness error in the second-order dispersion energy may not be negligible in this basis.[22]

IV. Which Model?

With the current version of the CamCASP program, distributed nonlocal polarizabilities can be calculated up to rank 4 using the constrained density-fitting algorithm.[8] If the molecule contains n_s sites and $l_{\max}$ is the highest rank included, the nonlocal polarizability model contains $O(n_s^2 l_{\max}^2)$ polarizability components. This can be many thousands for a rank 4 description of a molecule like BOQQUT, which contains 22 atoms. The computational cost of using such a polarizability model is already quite high for a pair of interacting molecules. For a cluster of molecules, a model of such complexity could be impossible to use. However,

Table 1. Maximum and rms Differences between the Model and SAPT(DFT) Second-Order Induction Energies of the Molecule Interacting with a Unit Charge on the vdW × 2 Surface[a]

	formamide		N-MPA		benzene		BOQQUT	
model	max	rms	max.	rms	max	rms	max	rms
NL1	5.40	2.79	7.63	3.95	5.35	3.84	7.35	4.93
NL2	1.48	0.48	1.42	0.57	1.60	0.66	1.92	0.78
NL3	0.57	0.27	0.36	0.17	0.30	0.19	0.54	0.15
NL4	0.56	0.26	0.37	0.17	0.27	0.17	0.57	0.14
L1	3.93	1.90	5.56	2.30	3.34	1.98	5.08	2.35
L2	6.85	1.37	4.66	2.02	3.71	1.74	7.90	1.75
L1,WSM	3.48	1.35	2.32	0.74	2.98	1.58	2.48	0.82
L2,WSM	1.14	0.27	1.29	0.21	0.69	0.25	2.27	0.23
L2/L1,WSM	1.18	0.34	0.99	0.34	0.67	0.42	1.48	0.32

[a] 'NL*n*' denotes a nonlocal model of rank *n*, 'L*n*' denotes a local model of rank *n* obtained using the Le Sueur and Stone[23] localization method, and 'L*n*,WSM' denotes a refined, local model of rank *n* obtained using the WSM procedure. The mixed-rank descriptions are denoted by 'L2/L1'. This means we have used a rank 2 local description on the heavy atoms and a rank 1 local description on the hydrogen atoms. All differences are reported in kJ mol^{-1}.

dramatic simplifications in the polarizability model can be made without significant losses in accuracy, using the localization techniques described in section IV.3 of part 1.

In Table 1 we report the maximum and rms errors made by several polarizability models in reproducing the induction energies of a molecule with a unit point charge placed on the vdW × 2 surface (as described in section II). Notice that the exchange part is zero in this case, because the point charge carries no electrons, so we are studying the ability of distributed polarizability models to reproduce $E_{\mathrm{ind,pol}}$. The exchange part is not negligible in general, but it is short-range in form and cannot in any case be described by a classical polarizability model.

The rank 1 nonlocal models are clearly inadequate, with maximum and rms errors in the range 3−8 kJ mol^{-1} or, for a more realistic charge of 0.5*e*, between 1 and 2 kJ mol^{-1}. These errors are dramatically reduced in the rank 2 nonlocal models and are still better for the rank 3 models for which the errors are only a few tenths of a kJ mol^{-1}. The rank 4 nonlocal models offer negligible improvements over the rank 3 models.

Curiously, at rank 1, the Le Sueur and Stone[23] localization technique results in local models that are *more accurate* than the nonlocal models they were constructed from. At present we do not understand why this is so. It is surprising because, as has been mentioned earlier, the Le Sueur and Stone localization procedure uses truncated multipole expansions to transform the nonlocal terms away, a procedure which is expected to cause a deterioration in the convergence properties of the model by increasing its sphere of divergence.

At rank 2 we see that the Le Sueur and Stone procedure does indeed lead to a deterioration compared with the nonlocal models. While the rms errors are slightly improved over the rank 1 local models, with the exception of *N*-methyl propanamide, the maximum errors are much larger, being in the range 4−8 kJ mol^{-1}, or, for a charge of 0.5*e*, 1−2 kJ mol^{-1}. These errors are probably too large for most applications.

We can obtain more accurate local models by refining the results of the Le Sueur and Stone localization procedure, as described in section IV.3 of part 1. For this refinement, which is the last stage in the WSM procedure, we need to choose the coefficients $g_{kk'}$, in eq 36 in part 1, which determine the weight given to the 'anchors' of the polarizability values. We have found that the following values lead to accurate models while minimizing the occurrence of unphysical terms:

$$g_{kk'} = 0 \text{ if } k \neq k'$$

$$g_{kk} = \begin{Bmatrix} 10^{-5} & \text{if } k \in \{10, 10c, 10s\} \\ 0 & \text{otherwise} \end{Bmatrix} \qquad (2)$$

The dipole−dipole polarizabilities from the constrained density-fitting and Le Sueur and Stone procedure are usually quite accurate, so they are used as anchor values and given nonzero weight, while the higher ranking polarizabilities can be quite poor and are given zero weight in the WSM procedure. The accuracy of the polarizability descriptions obtained from the constrained-DF method varies considerably with molecule size, so a system-dependent weight might yield better results. While the above choice of the weights has been found to be appropriate for all the molecules studied in this work, it is possible that there will be exceptions which will require another choice. In general, the errors made by the models in reproducing the point-to-point polarizabilities or the induction energy with a point charge (see below) should be monitored. As a rule of thumb, the maximum and rms errors in the point-to-point polarizabilities as percentages of the range[8] should be less than 6% and 0.2%, respectively, for a rank 1 model and less than 2% and 0.05% for a rank 2 model.

From Table 1 we see that the WSM rank 1 local models are already rather good, with maximum and rms errors of around 3 and 1 kJ mol^{-1}, respectively, for unit probe charge. At rank 2, the WSM models are good for all molecules. In all cases, the rms errors are only a few tenths of a kJ mol^{-1}. Maximum errors are somewhat larger at around 1−2 kJ mol^{-1}. These are all for unit charge and would be smaller by a factor of 4 or so for a more realistic charge.

In Figure 5 we show difference maps of the induction energy of the formamide molecule in the field of a unit point charge, computed using the nonlocal models, local models, and WSM local models, respectively. The large errors in the nonlocal rank 1 model, particularly near the oxygen and the polar hydrogen atoms, are quite clearly displayed. These errors are reduced in the rank 1 local model and are still smaller in the WSM rank 1 local model. The largest residual errors always seem to occur in the same regions, perhaps indicating the need for higher ranking polarizabilities on some sites than on others. At rank 2, the nonlocal model is in almost perfect agreement with SAPT(DFT), but the local model obtained using the Le Sueur and Stone localization method exhibits rather large deficiencies near the polar hydrogens. These are removed in the WSM rank 2 local model which is comparable in accuracy to the rank 2 nonlocal model.

By and large, the behavior of the polarizability models for *N*-methyl propanamide, benzene, and 3-azabicyclo[3.3.1]-

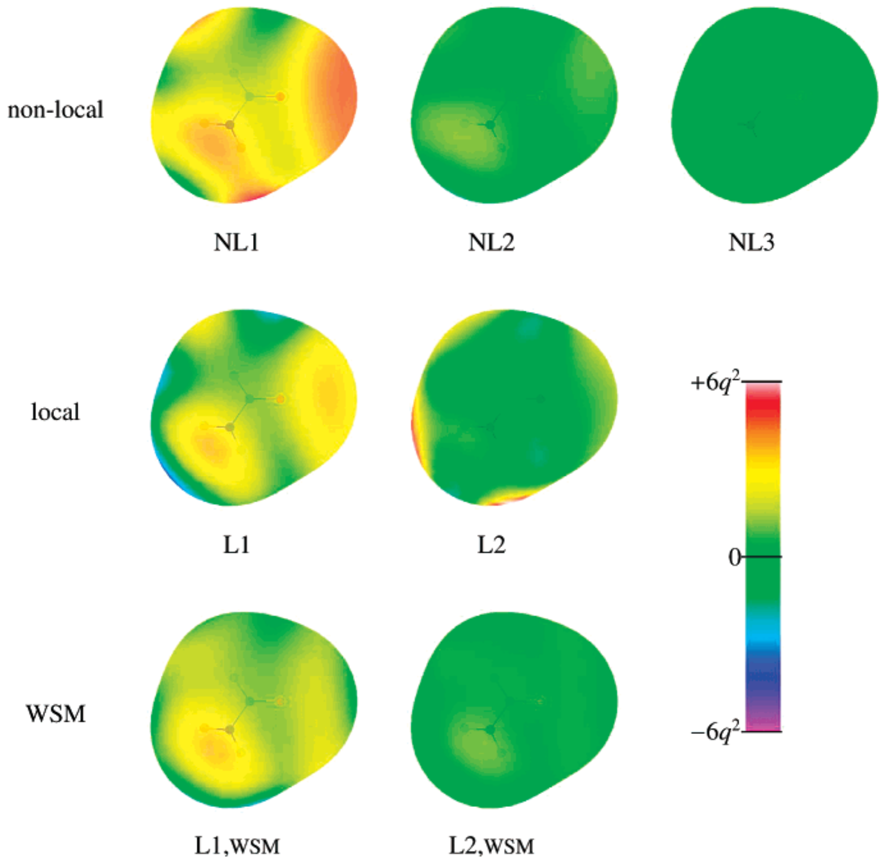


Figure 5. Difference maps of the induction energy (kJ mol^{-1}) arising from a charge q atomic units on the vdW × 2 surface of formamide using distributed nonlocal description of ranks 1, 2 and 3, distributed local description of ranks 1 and 2 obtained from the nonlocal models using the Le Sueur and Stone localization technique,[23] and WSM distributed local descriptions of ranks 1 and 2. The differences are taken against SAPT(DFT) second-order induction energies obtained using a molecular description with the Sadlej/MC basis set.

nonane-2,4-dione (BOQQUT) are similar to those for formamide, so we will comment only briefly on the models for these systems. The SAPT(DFT) induction energy maps and difference maps for the WSM rank 1 and rank 2 local models for these molecules are displayed in Figure 6. It is clear that the WSM rank 1 local models consistently underestimate the induction energy arising from a point charge, while the WSM rank 2 models offer consistently higher accuracy. This underestimation is notably severe for the benzene molecule, because the π-orbitals need to be described by rank 2 polarizability terms.

More accurate models are possible, but only if we are willing to accept the presence of unphysical terms. Even in the present models, some of the quadrupole−quadrupole polarizabilities violate the requirement of positive-definiteness, but, in the absence of any constraints, even the dipole−dipole terms are not always positive-definite.

The distributed polarizability calculations on BOQQUT brought to light a potential problem with the WSM procedure as currently implemented. BOQQUT is a fairly large molecule and possesses a plane of symmetry. We first calculated refined polarizabilities for this molecule using point-to-point polarizabilities calculated on a grid of 1000 points. The resulting rank 2 local model significantly broke the symmetry of the molecule, because the grid of points used for calculating the point-to-point polarizabilities was chosen at random and did not respect the symmetry of the molecule. The asymmetry could have been avoided simply by ensuring that the parameters of the polarizability model were correctly symmetrized, and we normally do this, but the use of unsymmetrized parameters allowed us to assess the quality of the grid, by determining how large the grid needed to be before the asymmetry was numerically negligible. For BOQQUT, a grid of 2000 points, or over 2 million point-to-point polarizabilities, proved to be adequate. The calculation of the point-to-point polarizabilities needed less than 2 h of CPU time on a single processor. This is due to the computational efficiency of the density-fitting-based algorithm[8] that has been implemented in the CamCASP program[15] and used for the point-to-point polarizabilities.

V. Penetration, Truncation Errors, and Damping

The multipole expansion provides us with a computationally efficient means of calculating the induction energy. However the resulting energies will be in error for a number of reasons, which must be addressed if we are to ensure accurate interaction energies.

(1) The multipole series are expansions in inverse powers of the intersite distance R_{ab}, so they diverge when the sites coincide. The multipoles and polarizabilities that appear in the damped classical polarizable model (eqs 21 and 22 in

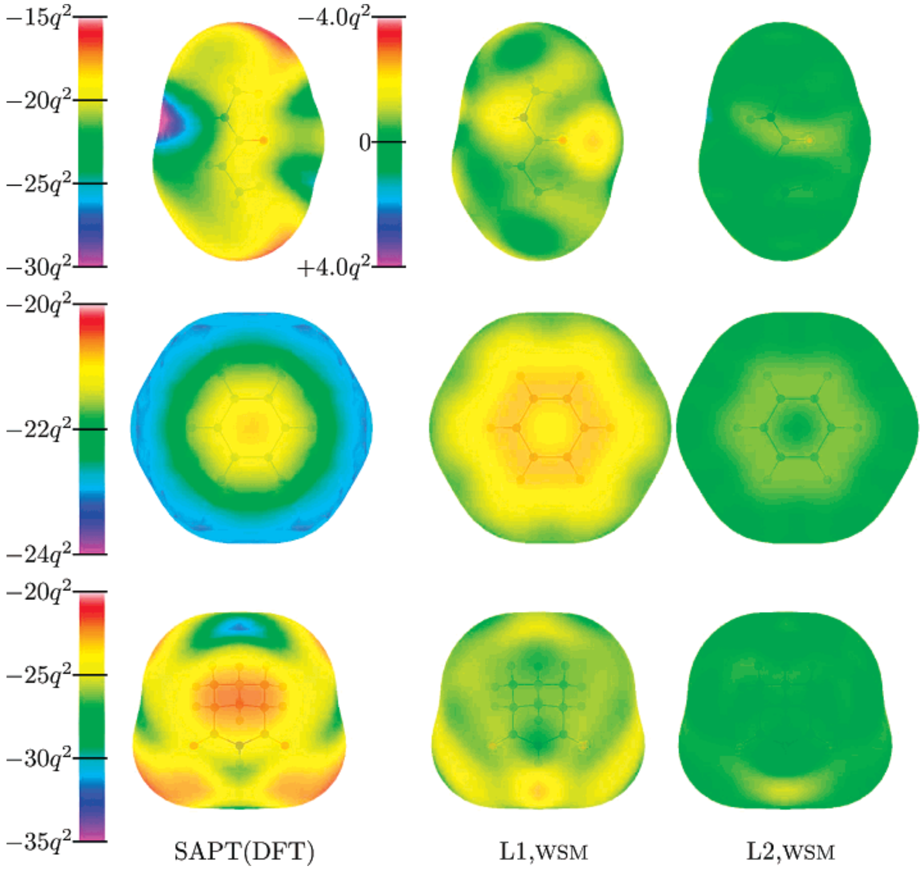


Figure 6. Induction energy maps and difference maps (kJ mol^{-1}) with the Sadlej basis arising from a charge q atomic units on the vdW × 2 surfaces of the *N*-methyl propanamide (top row), benzene (middle), and BOQQUT (bottom) molecules. The induction energy maps have been obtained using SAPT(DFT). The difference maps for the WSM rank 1 and rank 2 local polarizability models have been taken against the corresponding SAPT(DFT) second-order induction energies obtained using molecular descriptions with the Sadlej/MC basis set. The scale used for the difference maps is the same for all three molecules and is shown for *N*-methyl propanamide only.

part 1) treat the electron charge distribution as if it were concentrated at the local origin. That is, the finite extent of the charge distribution is neglected and orbital penetration effects are absent. The exchange part of the induction energy is also absent from the multipole expansion. These features result in a ‘penetration error’. The true damping function—if one could be found—would account for this error.

For the second-order two-body induction energy, the divergence occurs at small enough intersite distances that it can be ignored, except for Monte Carlo simulations where a trial step is taken without regard to energy. Nevertheless, when iterations are included in the damped classical polarizable model (part 1, section IV.2) the role of the damping becomes more important. This can be seen from eq 22 in part 1, for at the *m*th iteration, the expression involves the product of $m + 1$ damping functions. The number of iterations needed for convergence increases as the intersite distance decreases, so the cumulative effect of the damping increases.

(2) Additionally, in practice, the multipole expansion must be truncated at a low rank. This introduces a ‘truncation error’.

These sources of error need to be accounted for in accurate calculations of the interaction energies. This can be done using comparisons with the nonexpanded energies. While this comparison is quite straightforward for the electrostatic energy,[7] it is much less straightforward for the induction energy, because it is not immediately clear which SAPT-(DFT) energies we should be using as a reference.

It has generally been assumed that the expanded induction energies approximate the induction energy from the polarization approximation. After all, the former is derived from the latter by using the multipole-expanded form of the interaction operator.[24] For example, in the absence of iterations, the expanded induction energies, $E^{(2)}_{\text{ind,d-class}}$, have been expected to approximate $E^{(2)}_{\text{ind,pol}}$. From the discussion in section IV.1 of part 1 and section III, it should come as no surprise that this is generally not the case. $E^{(2)}_{\text{ind,d-class}}$ does indeed approximate $E^{(2)}_{\text{ind,pol}}$ rather well if medium-sized monomer basis sets are used.[7] This is because the spurious tunneling effects discussed in section IV.1 of part 1 and, consequently, the exchange-induction energies are small for such basis sets. However, when large monomer basis sets are used, and especially with the MC$^+$ basis type, the spurious tunneling effects are quite large and result in large (negative) values for $E_{\text{ind,pol}}$. In such a case, $E_{\text{ind,exch}}$ is also large (and positive) and quenches $E_{\text{ind,pol}}$ quite substantially. These tunneling effects are completely absent from the expanded form of the induction energy based on eq 21 in

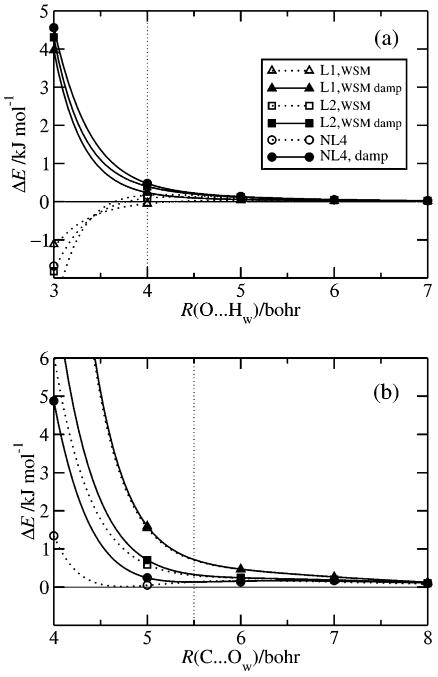


Figure 7. Errors in the expanded induction energy for the formamide−water dimer. The energy difference plotted on the *y*-axis is defined as $\Delta E = E^{(2)}_{\mathrm{ind,d-class}} - E^{(2)}_{\mathrm{ind,tot}}$, so a positive ΔE means that the expanded energies are not attractive enough. The SAPT(DFT) $E^{(2)}_{\mathrm{ind,tot}}$ energies have been calculated with a Sadlej/MC$^+$ basis. Results are displayed for two relative orientations of the formamide and water molecules as described in Figure 3. The dotted vertical lines mark the approximate minimum-total-energy separations. Notice that the damped and undamped rank 1 curves for geometry (b) are almost identical. All local polarizability models have been obtained using the WSM procedure.

part 1. At present we cannot tell whether it is more appropriate to compare $E^{(2)}_{\mathrm{ind,d-class}}$ with $E^{(2)}_{\mathrm{ind,pol}}$, calculated in the monomer basis set, or with $E^{(2)}_{\mathrm{ind,tot}}$, that is, the sum of $E^{(2)}_{\mathrm{ind,pol}}$ and $E^{(2)}_{\mathrm{ind,exch}}$, calculated with the MC$^+$ basis set. The former will include some spurious effects; in the latter they will be larger but partly cancelled out. We are currently investigating this question, but for the present work we use the latter choice in the comparisons of the expanded models below.

We now illustrate the points discussed above with the noniterated induction energies, i.e., $E^{(2)}_{\mathrm{ind,d-class}}$, calculated for the formamide−water dimer using the Sadlej/MC$^+$ basis set. Figure 7 shows the energy difference $\Delta E = E^{(2)}_{\mathrm{ind,d-class}} - E^{(2)}_{\mathrm{ind,tot}}$. The two dimer geometries used, though somewhat artifical, serve the purpose of representing the two main types of intermolecular interactions: geometry (a) is a hydrogen-bond-like interaction with the hydrogen on water making close contact with the oxygen on formamide, and geometry (b) is a nonpolar interaction with the oxygen on water in close contact with the carbon on formamide. First we focus on the results without damping. The rank 4 nonlocal model provides an excellent description for both geometries, with energy differences less than 0.2 kJ mol^{-1} for physically relevant intersite separations. The rank 3 nonlocal model (not shown) gives very similar results to the rank 4 model, which suggests that increasing the rank above 4 will not improve the description significantly. Therefore the residual difference must be due to orbital penetration (a positive energy difference) and the divergence of the multipole expansion at short-range (a negative energy difference). Being of opposite signs, these two effects partially cancel. For geometry (a), the relatively small charge density on the hydrogen atom and the small intersite distance means that the divergence of the multipole expansion is the dominant effect and the rank 4 nonlocal model needs net damping; but for geometry (b), the opposite is the case, and the rank 4 nonlocal model needs a net enhancement at short intersite separations.

The rank 1 and 2 local models behave in a similar manner. However, while all three models result in similar energies for geometry (a), the rank 2 model is substantially better than rank 1 for geometry (b). This probably indicates the importance of higher ranking terms in the polarizability description of the heavy atoms.

We now turn to the issue of damping. As has been mentioned above, the damping function has to account for penetration effects and eliminate the divergence of the multipole expansion that occurs at small intersite distances. These two effects are of opposite signs and very likely depend differently on the intersite separation, as is the case for geometry (b) in the formamide water example. In practical applications the damping function must also compensate for truncation effects, which can complicate the picture quite substantially, as we have seen from the above example. It is unrealistic to expect to find a universal damping function capable of accounting for all these effects. The damping function will have to depend on the sites involved and will probably have to be larger than unity at intermediate distances, where enhancement is needed to account for truncation and penetration effects, and less than unity at short distances, to cancel out the short-range divergence of the multipole expansion. Such a function has been proposed for the dispersion energy,[25] but to the best of our knowledge, not for the induction energy.

We believe that the best we can do at present is to damp out the short-range divergence of the multipole expansion. This is especially necessary in calculations of the induction energy of clusters of polar molecules. Intersite distances can be quite small in such clusters, because cooperative induction effects can be quite large and compensate for the unfavorable exchange energies associated with small intersite distances. In such cases, an undamped multipole expansion will result in nonsensical induction energies, particularly when iterations are included in the evaluation of eqs 21 and 22 in part 1. A convenient choice for the damping functions are those due to Tang and Toennies[26] which have had the greatest success in the generation of high-accuracy potentials for small dimers. This will leave a residual error arising from truncation and penetration effects which can, in principle, be accounted for using an overlap model.[7]

We are then led to the problem of determining the damping factor to be used in the damping functions. It is often assumed that the damping factor can be determined by a comparison of the expanded and nonexpanded energies.

Given the observations of the above paragraphs, such a comparison is fruitless as it would suggest an antidamping. An alternative procedure has been to use the exponential parameter in the Born−Mayer term from a fit to the exchange energies as the damping parameter. This could, in principle, be done on a site−site basis to obtain damping parameters that depend on the pair of sites involved. The argument for this approach is that the exchange−repulsion and damping are both consequences of the wave function overlap. Perhaps a simpler method of determining the damping factor β, though one that depends only on the interacting molecules and not on the individual interacting sites, is based on the considerations presented in Chapter 6 of ref 24 and goes as follows: The potential due to the electronic density of a hydrogen-like atom with wave function $\psi(r) = \sqrt{\alpha^3/\pi}\, e^{-\alpha r}$ is

$$V(r) = -\frac{1}{r} + e^{-2\alpha r}\left(\alpha + \frac{1}{r}\right) \qquad (3)$$

The first term in this equation is the multipole expansion of a spherical electron cloud and the second is the penetration correction. This potential can be rewritten as

$$V(r) = -\frac{1}{r} f_1(2\alpha r), \quad \text{where } f_1(2\alpha r) = 1 - e^{-2\alpha r}(1 + \alpha r) \quad (4)$$

so that f_1 is the damping function that correctly incorporates the penetration effect. This result can be plausibly, though not rigorously, generalized to an arbitrary wave function by using the asymptotic form of the wave function[27]

$$\psi(r) \rightarrow e^{-\sqrt{2I}r} \qquad (5)$$

where I is the vertical ionization potential. We now see that the damping factor should be

$$\beta = 2\sqrt{2I} \qquad (6)$$

where everything is in atomic units, so β is in bohr^{-1} if I is in Hartree. For most organic molecules, this results in an atom−atom damping function with a damping constant between 1.9 and 1.7. For mixed dimers, we suggest using the damping factor $\beta = \sqrt{2I_A} + \sqrt{2I_B}$. This choice is plausible as it is the coefficient of R in the exponential factor of the density-overlap function.

There is reason to believe this is a good choice for the damping factor, but there may be cases for which the issue of damping will have to be re-examined. Numerical evidence from calculations of the induction energy of organic crystals[28] indicates that above damping factor is not only appropriate but also essential when the rank 2 models needed for high accuracies are used. Welch et al.[28] have observed that without damping the small intersite distances present in some organic crystals can lead to very large and unphysical crystal induction energies. However, using the above damping factor leads to a rational progression of the crystal induction energy with rank of the polarizability description. As would be expected, damping has the largest effect on the higher ranking models.

However, there will be molecules for which a single damping factor may be too simplistic. Our definition of β involves the ionization potentials of the molecules. For a large molecule, with very different functional groups, it may be necessary to use a different ionization potential for each of the functional groups. This would then lead to a damping factor that depended on the pair of interacting groups. We have not yet investigated such a possibility.

In the formamide−water example, using ionization potentials of 0.375 au and 0.464 au for formamide and water, respectively, we obtain $\beta = 1.83$. From Figure 7 we see that after damping all models exhibit a positive energy difference which is a combination of penetration effects and truncation errors. From the modeling point of view, their uniformity makes these differences simpler to handle than those made by the undamped models. For example, as suggested in ref 7, the overlap model[29] could be used to model the residual energy differences. We are currently working on a methodology to make this possible on a routine basis.

VI. Approximations

The many issues discussed in the preceding sections are important in high-accuracy calculations of the interaction energy. However, for many applications, it may be sufficient to achieve a moderate accuracy. Indeed, for large systems, some of the prescriptions given above may not even be feasible due to lack of computational resources. Here we discuss the approximations most useful in calculations on large systems and test their accuracy.

VI.1. Calculating $E^{(2)}_{\text{ind,tot}}$ and $E^{(3)}_{\text{ind,tot}}$ Using a Monomer Basis. In section III we have argued that the ‘far-bond’ functions that are part of the MC$^+$ type of basis should be used in order to obtain basis-saturated induction energies. Since the far-bond functions are placed at the locations of the nuclei of the interacting partner, this means that the MC$^+$ basis set depends on the dimer geometry. This in turn means that the calculation of the Hessians (eqs 7 and 8 of part 1), which is the most computationally expensive step in the evaluation of the induction energy, needs to be repeated for each dimer geometry. This can be computationally prohibitive and would be avoided if the monomer basis (MC) were to be used. Furthermore, there is good evidence to suggest that the MC$^+$ basis type introduces significant errors in the first-order energies[30] for which the MC type should be used.

The first-order energies are reasonably well converged with the Sadlej/MC basis set. The problems lie with the second-order energies. Both the induction and dispersion energies are insufficiently converged in this basis. As we saw in section III, the so-called mid-bond basis functions that are included in the MC$^+$ basis type are needed to obtain basis-converged dispersion energies. Like the far-bond functions, the mid-bond basis functions also make the basis set dependent on the dimer geometry. Neglecting the mid-bond functions will introduce significant errors in the dispersion energy which will complicate the discussion of the induction energy calculations. Therefore, for the purposes of the present paper, we will avoid discussing the basis convergence issues associated with the dispersion energy by using dispersion

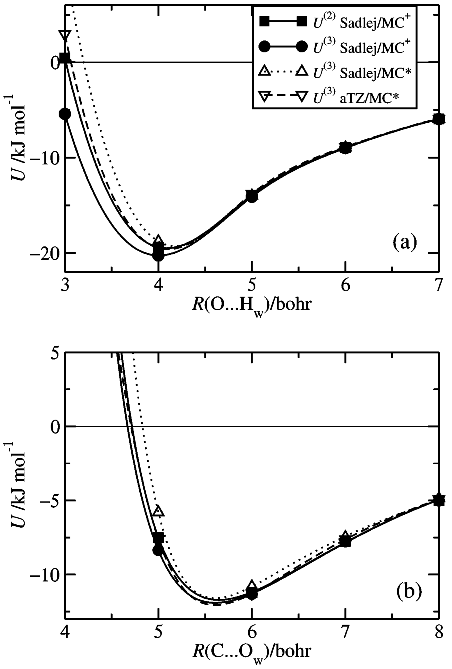


Figure 8. Total interaction energies for the formamide−water dimer using different basis sets. $U^{(2)}$and $U^{(3)}$ are the total SAPT(DFT) interaction energies up to second and third order, respectively. Calculations have been performed using the MC$^+$ and MC basis types. The latter basis has been labeled as 'MC*' because, for reasons explained in the text, the dispersion and exchange-dispersion energies have been calculated using the MC$^+$ basis. The $U^{(2)}$ and $U^{(3)}$ potential curves with the MC* basis are very similar, so, for clarity, only the latter are displayed. Results are displayed for two relative orientations of the formamide and water molecules as described in Figure 3. For geometry (b), the $U^{(3)}$ aTZ/MC* curve is almost obscured by the $U^{(2)}$ Sadlej/MC$^+$ and $U^{(3)}$ Sadlej/MC$^+$ curves.

energies computed using the Sadlej/MC$^+$ basis. We will denote the basis types for these mixed basis calculations by 'MC*'.

In Figure 8 we display total interaction energies for the formamide−water dimer obtained using different basis sets. For the MC* basis type only the $U^{(3)}$ potentials are shown, since $U^{(2)} \approx U^{(3)}$ for this basis type. This is because $E^{(3)}_{\text{ind,tot}}$ is negligibly small when computed using monomer basis sets, as should be apparent from Figure 4. $U^{(3)}$ from the Sadlej/MC* basis results in potentials that are consistently too shallow, particularly for the hydrogen-bonded geometry (a). A considerable improvement is obtained if the aug-cc-pVTZ/MC* basis is used. For both geometries, this basis gives potential curves in good agreement with the $U^{(2)}$ potential obtained with the Sadlej/MC$^+$ basis set. Therefore the aug-cc-pVTZ/MC* basis is a viable alternative to the dimer-geometry dependent Sadlej/MC$^+$ basis.

It must be borne in mind that the aug-cc-pVTZ/MCbasis is twice the size of the Sadlej/MC basis and larger even than the Sadlej/MC$^+$ basis. Therefore, while the aug-cc-pVTZ/MC basis allows fairly accurate induction energies to be evaluated for all dimer geometries using Hessians calculated just once, it does entail a significant increase in the computational cost of evaluating the Hessians.

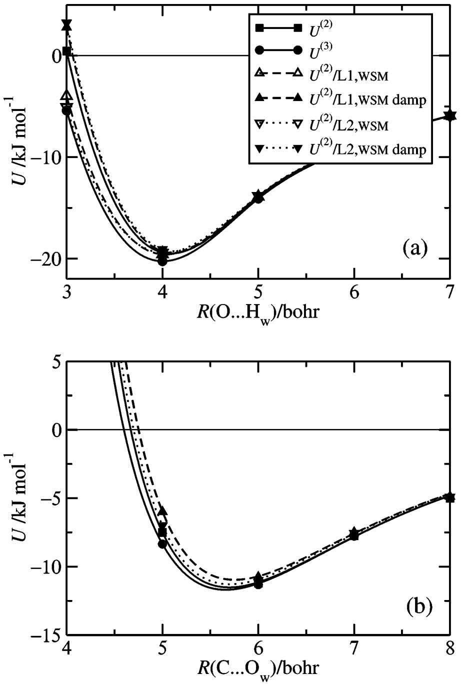


Figure 9. Total interaction energies for the formamide−water dimer. $U^{(2)}$ and $U^{(3)}$ are the SAPT(DFT) interaction energies at second and third order. Approximations to $U^{(2)}$ are labeled $U^{(2)}$/L*n*,WSM,(damp). This notation means that the SAPT(DFT) induction energies have been replaced by the induction energies obtained from a damped classical polarizable model with iteration, using a distributed polarizability model of rank *n* on both molecules. Results are displayed for two relative orientations of the formamide and water molecules as described in Figure 3. For geometry (a), the $U^{(2)}$/L*n*,WSM and $U^{(2)}$/L*n*,WSM,damp curves are just above the $U^{(3)}$ and $U^{(2)}$ curves, respectively. For geometry (b), damping has almost no effect on the energies, so the damped and undamped curves are indistinguishable. All local polarizability models have been obtained using the WSM procedure.

VI.2. Neglecting the Higher-Order Two-Body Energies. The calculation of the higher-order energies using eq 25 in part 1 is the most time-consuming part of a SAPT(KS) calculation. Its computational scaling is dominated by the evaluation of $E^{(3)}_{\text{ind,exch}}$[KS] which scales as $O(n_o^2 n_v^3)$, where n_o and n_v are the number of occupied and virtual orbitals, respectively. From the discussion in section IV.1.2 of part 1 we know that the higher-order energies cannot be neglected in systems that exhibit strong hydrogen bonds, such as water and hydrogen fluoride, but their effect is far less in other systems and may even be negligible. In Figure 8 we display total interaction energies for the formamide−water system in the two representative geometries used in earlier discussions. In geometry (a), the hydrogen-bonded geometry, the $U^{(3)}$ potential curve is deeper than $U^{(2)}$. Using the $U^{(2)}$ potential curve would result in a slightly longer formamide−water bond, but the errors are not large and could well be acceptable in calculations of moderate accuracy.

The situation is far better for the non-hydrogen-bonded geometry (b). Here the $U^{(3)}$ potential curve is only slightly deeper than $U^{(2)}$, which could be used with almost no loss in accuracy.

Table 2. Contribution of Third- and Higher-Order Corrections to the Interaction Energy for the Water, Hydrogen Fluoride, Carbon Dioxide, and Benzene Dimers and the $H_2O\cdots H_3N$ and $H_2\cdots CO$ Complexes[a]

	$(H_2O)_2$	$(HF)_2$	$H_2\cdots H_3N$	$(CO_2)_2$	$(C_6H_6)_2$	$H_2\cdots CO$
$E^{(2)}_{\mathrm{ind}}$ SAPT(DFT)	−5.164	−6.687	−0.479	−0.866	−0.885	−0.122
$E^{(2)}_{\mathrm{ind,tot}}$/L1,WSM,damp	−3.932	−4.268	−0.361	−0.358	−0.599	−0.073
$E^{(2)}_{\mathrm{ind,tot}}$/L2,WSM,damp	−4.507	−5.303	−0.357	−0.474	−0.710	−0.080
$E^{(2)}_{\mathrm{ind,tot}}$/NL4,damp	−5.051	−5.647	−0.420	−0.619	−0.754	−0.081

[a] The first three dimers are at their equilibrium geometries, and the benzene dimer is in the parallel stacked geometry with a center-of-mass separation of 3.8 Å. $H_2\cdots CO$ is in the linear geometry with C toward H_2 and a center-of-mass separation of 7.8 au, and $H_2O\cdots H_3N$ is in geometry (a) of Figure 5 from ref 31. L1,WSM and L2,WSM denote damped, noniterated, classical induction energies evaluated using local rank-1 and rank-2 polarizabilities refined using the WSM procedure, and NL4 denotes damped energies obtained from nonlocal rank-4 polarizabilities. Molecular properties needed for the damped classical model were obtained using the aug-cc-pVTZ/MC basis. All energies are reported in kJ mol^{-1}.

VI.3. Replacing$\boldsymbol{E^{(2)}_{\mathrm{ind,tot}}}$ and $\boldsymbol{E^{(3)}_{\mathrm{ind,tot}}}$ by $\boldsymbol{E_{\mathrm{ind,d-class}}}$. An approximation that is commonly used is to avoid calculating the nonexpanded induction energies altogether by using the expanded induction instead. From Figure 7 it should be apparent that this approximation would be rather good if the undamped rank 4 nonlocal polarizability description were used on both molecules, but this description is very elaborate and would probably be impossible to use in clusters of molecules. The rank 1 and rank 2 local polarizability descriptions are more practicable but less accurate. The accuracy can be improved slightly by using the iterated form of the polarizability models.

We will use $U^{(2)}$/Ln,WSM,(damp) to denote the total interaction energy obtained by replacing $E^{(2)}_{\mathrm{ind,tot}}$ and $E^{(3)}_{\mathrm{ind,tot}}$ with $E_{\mathrm{ind,d-class}}$, where n denotes the rank of the (possibly damped) local polarizability model. In Figure 9 we display total interaction energies obtained using these approximations. For geometry (a), both $E^{(2)}_{\mathrm{ind,tot}}$/L1,WSM,damp and $E^{(2)}_{\mathrm{ind,tot}}$/L2,WSM,damp are reasonably good approximations to $U^{(2)}$. Curiously, the undamped approximations are very close to $U^{(3)}$. For geometry (b), all approximations fare well, with the rank 2 models being the more accurate. Damping has a very small effect on the energies here. Therefore, $U^{(2)}$/L2,WSM,damp is a viable approximation to the total interaction energy in both cases.

Table 2 shows results for some other small dimers at equilibrium or near-contact geometries. Here too the results are generally satisfactory.

The computational cost of evaluating the induction energy using the rank 1 or 2 local polarizability models is much lower than the cost of evaluating $E^{(2)}_{\mathrm{ind,pol}}$ and $E^{(2)}_{\mathrm{ind,exch}}$ within SAPT(DFT). This is true even if the SAPT(DFT) energies are calculated using the MC type of basis. Additionally, we see from Figures 8 and 9 that $U^{(2)}$/Ln,WSM,(damp) is a better approximation to the interaction energy than calculating $U^{(2)}$ in the Sadlej/MC* basis. This is particularly welcome for applications involving large molecules for which we may be able to calculate molecular properties but not the SAPT(DFT) interaction energies. The relatively good accuracy of the $U^{(2)}$/Ln,WSM,(damp) approximation also helps explain why ab initio intermolecular potentials that used multipole expansion for the induction energy, such as the ASP water potential,[32] have been so successful.

VI.4. Mixed-Rank Polarizability Descriptions. There will be situations where polarizability models of mixed rank can be used. For example, an accurate description of the polarizability of benzene can be obtained with rank 1 polarizabilities on the hydrogen atoms and rank 2 terms on the carbon atoms.[9] For large, compact molecules, it may be desirable to simplify the polarizability model by reducing the rank of the description of those atoms hidden under the van der Waals spheres of neighboring atoms or even omitting them altogether from the model. Further simplifications can be achieved by enforcing symmetries of functional groups or eliminating small terms in the polarizability model. All of these simplifications can be incorporated in the WSM procedure described in section IV.3 of part 1. This procedure, together with the constrained density-fitting distribution method, gives us the flexibility to choose an appropriate polarizability description while avoiding unphysical terms in the polarizability description as far as possible.

In Table 1 we report the maximum and rms errors in polarizability models using a rank 2 description on the heavy atoms and a rank 1 description on the hydrogens. The rms errors in these mixed-rank polarizability descriptions are comparable to those made by the more complex rank 2 descriptions, while the maximum errors, which tend to occur near the hydrogen atoms, are significantly smaller for the larger molecules studied here. Additionally, there is far less loss of positive-definiteness in the mixed-rank descriptions as it is the quadrupole−quadrupole polarizability terms on the hydrogen atoms that tend to be negative.

In the case of BOQQUT, the mixed-rank description has nearly half as many nonzero polarizability components as the rank 2 description. The mixed-rank description of the formamide molecule results in induction energies of the formamide water dimer that are almost identical to those from the rank 2 local description.

We therefore have good physical and computational reasons for using the mixed-rank polarizability descriptions and strongly recommend use of these models in favor of the more complex and less physical rank 2 models.

VII. Summary

We have provided a theoretical and numerical framework for the accurate calculation of the induction energies of clusters of organic molecules. These are large systems and pose quite different problems from those that have faced the high-accuracy, small-molecule community. These problems can be broadly classified as those concerned with the theoretical details of the induction energies at second and higher order in the interaction operator and those concerned with numerical details and the development of models suitable for applications.

VII.1. Model Induction Energies. Accurate molecular polarizabilities and multipole moments are needed for modeling of the induction energy of clusters of molecules. For all but the smallest of molecules, these properties need to be distributed. The problem of distributing the multipole moments has already been addressed[33,34] in a satisfactory manner. Here we have proposed and demonstrated an accurate and versatile method of obtaining distributed polarizabilities that is suitable for molecules of as many as 30 atoms or so. Our distribution scheme for the polarizabilities is based on the methods of Williams and Stone[9] and our constrained density-fitting method.[8] By combining the strengths of these two methods, we have obtained a distribution procedure with properties that make it ideal for high-accuracy calculations on systems of large molecules. The main features of this distribution scheme are as follows:

(1) The underlying theory used in the polarizability calculations is coupled Kohn−Sham theory (CKS), also known as Kohn−Sham linear response theory. Molecular properties obtained using CKS theory can exceed coupled-cluster methods in accuracy when used with a modern density functional like PBE0[35] with asymptotic corrections,[36,37] thus ensuring an accurate polarizability description at modest computational cost. In the form used in this paper and implemented in the CamCASP program,[15] the CKS equations are solved with a computational effort that scales as $O(n_o^3 n_v^3)$, but this scaling can be reduced using density-fitting techniques.[5,38,39] With the current implementation of the CKS equations we have been able to compute the properties of the 3-azabicyclo[3.3.1]nonane-2,4-dione (BOQQUT) molecule, containing 22 atoms, in less than a day of CPU time on a single Opteron processor using the Sadlej basis set. With the density-fitted form of these equations we expect to be able to perform calculations on even larger systems.

(2) We have tested localized distributed polarizability models of rank 4 for small molecules and rank 2 or 3 for the larger molecules. In principle, local descriptions up to rank 4 could be generated for all molecules, but rank 2 should be sufficient.

(3) Non-positive-definite polarizability tensors do not occur in the dipole−dipole polarizabilities, and at higher rank they arise mainly for the hydrogen atoms. Consequently, this problem is smallest in the mixed-rank models.

(4) There is no fundamental limit to the size of the basis sets that can be used, though computational limitations will restrict it in practice.

(5) Finally, one of the most powerful features of the distribution scheme proposed here is that it gives us the ability to choose any reasonable polarizability model and yet obtain an accurate, physically correct, local polarizability model.

We have tested this Williams−Stone−Misquitta (WSM) distribution scheme using the formamide and *N*-methyl propanamide molecules and have also generated local polarizability models of ranks 1 and 2 for benzene, BOQQUT, and other molecules,[28] only some of which have been reported in this paper. The rank 2 local description is comparable in accuracy to the much more complex rank 4 nonlocal models. The rank 1 models underestimate the induction energy, particularly around the heavy atoms. Mixed rank models with a rank 2 description on the heavy atoms and rank 1 on the hydrogen atoms offer a good compromise between accuracy and simplicity.

The newly developed visualization techniques recently implemented in the ORIENT program[16] have given us a very powerful means of evaluating the polarizability models. By 3-dimensional visualization of the induction maps or error maps made against accurate SAPT(DFT) induction energies, we were able to make assessments of the shortcomings of these models. The 3-D maps enable us to identify sites at which a particular polarizability description may be deficient. This proves invaluable in designing accurate polarizability models where a mixed rank description may be necessary.

VII.2. Numerical Aspects. One of the main considerations in any ab initio calculation is the type of basis set to be used. It will not usually be possible to use large basis sets in calculations on organic molecules of the size considered in this paper, so we have attempted to determine which basis sets are good enough for calculations of the induction energy. The necessary level of accuracy will depend on the application, but for systems of organic molecules with dimer binding energies of 10−20 kJ mol^{-1}, basis set incompleteness errors of less than a few tenths of a kJ mol^{-1} at the important dimer geometries are probably acceptable.

From numerical tests on a variety of molecules (only one example was reported here), we recommend the Sadlej basis sets[18,19] for calculations of molecular multipole moments and polarizabilities. Auxiliary basis sets tuned for the Sadlej bases are not available, but we have found that the aug-cc-pVTZ auxiliary basis,[40,41] though probably too large, works very well. With these basis sets, very accurate polarizability models can be obtained. Comparisons with the much larger aug-cc-pVQZ basis show that the maximum error made by the Sadlej basis models is about 1.5 kJ mol^{-1} on the vdW × 2 surface in the field of a unit point charge. For a more realistic charge of 0.5 units, this would be an error of only 0.4 kJ mol^{-1}.

To get a similar accuracy, SAPT(DFT) induction energies must be evaluated using the Sadlej/MC$^+$ basis, that is, with the inclusion of basis functions located at the positions of the nuclei of the partner molecule.[20] The Sadlej/MC$^+$ and aug-cc-pVTZ/MC$^+$ bases give almost identical induction energies. However, when used in the MC basis type, that is, without the extra off-atomic functions, both the Sadlej and aug-cc-pVTZ bases yield poor induction energies, the latter being the better choice.

On the issue of damping: in our opinion, it is impossible for the damping functions in current use to recover the penetration energy. All that is possible is the damping out of the short-range divergence in the multipole expansion. Apart from special cases such as Monte Carlo simulations, this is not essential for the dimer energy, as intersite distances in a dimer are never small enough to see the onset of the divergence in the multipole series. However, in the bulk, many-body effects can cause small intersite distances, particularly in hydrogen-bonded geometries, and consequently damping is needed. We have argued that comparisons with nonexpanded energies are not useful in determining the damping needed. Rather, we propose that the Tang−Toennies functions[26] be used with a damping coefficient of $\beta = \sqrt{2I_A} + \sqrt{2I_B}$, where I_A and I_B are the ionization energies (in au) of the interacting molecules.

The remaining error will be entirely due to penetration effects and the truncation error, the latter arising from the truncation of the multipole ($1/R$) expansion. In accurate work, these errors must be accounted for in some way. As they decay exponentially with distance, like the exchange−repulsion energy, they could be included with it and modeled in a similar way. We are currently looking for a robust way to do this.

Comparisons of the model induction energies and the nonexpanded SAPT(DFT) energies brought to light a very unexpected correspondence. Contrary to supposition, the model induction energy at second order, $E^{(2)}_{\mathrm{ind,d-class}}$, does not recover the expression for the second-order induction energy in the polarization approximation, $E^{(2)}_{\mathrm{ind,pol}}$, but rather the *sum* $E^{(2)}_{\mathrm{ind,pol}} + E^{(2)}_{\mathrm{ind,exch}}$. We have defined this sum to be the induction energy, $E^{(2)}_{\mathrm{ind,tot}}$. The reason for this has to do with spurious tunneling effects present due to Coulomb singularities in the interaction operator.[13,14] These singularities are also responsible for the slow convergence of the induction energy with basis set and order in perturbation theory. It is expected that a SAPT(DFT) formulation with a regularized form of this operator will be free from both these problems. Preliminary investigations suggest that this is indeed the case.

We have tried to describe ways of calculating the induction energy with as few approximations as possible. However, there will always be systems for which approximations will be needed. Consequently, we have proposed and analyzed a number of possible approximations of varying complexity. The most useful of these approximations is one in which we calculate the induction energy of a cluster from the induction models only. We recommend using the damped mixed-rank local polarizability description, that is, with a rank 2 description on the heavy atoms and a rank 1 description on the hydrogen atoms. The success of this type of model also explains why potentials that have used a similar description of the induction energy have been so successful. An example is the ASP water potential.[32]

VIII. Programs

Many of the theoretical methods described in this review are implemented in programs available for download. Some of these, together with their main uses in the present work, are as follows:

(1) SAPT2002:[42] SAPT(KS) energy calculations.

(2) CamCASP 4.5:[15] Molecular properties in total and distributed form and SAPT(DFT) dispersion and induction energies. The CamCASP suite includes the GDMA 2.2 program[34] used for calculating the distributed multipoles needed in the induction energy calculations and the PFIT program used to refine the distributed polarizabilities in the WSMprocedure.

(3) ORIENT 4.6:[16] Localization of the distributed polarizabilities and visualization of the energy maps.

(4) DALTON 2.0:[43] DFT and CKS calculations. A patch[42] is needed to enable DALTON 2.0 to work with SAPT2002 and CamCASP 4.5.

Acknowledgment. We would like to thank Gareth Welch for useful discussions and comments and his help in obtaining the structure of BOQQUT. A.J.M. would like to thank Girton College, Cambridge for a research fellowship. This research was supported by EPSRC grant EP/C539109/1.

References

(1) Misquitta, A. J.; Stone, A. J. Accurate induction energies for small organic molecules: I. Theory. *J. Chem. Theory Comput.* **2007**, *3*, 7−18.

(2) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **2003**, *91*, 33201.

(3) Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.* **2005**, *122*, 214109.

(4) 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 theory. *J. Chem. Phys.* **2005**, *123*, 214103.

(5) 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.

(6) Misquitta, A. J.; Szalewicz, K. Intermolecular forces from asymptotically corrected density functional description of monomers. *Chem. Phys. Lett.* **2002**, *357*, 301−306.

(7) Stone, A. J.; Misquitta, A. J. Atom−atom potentials from *ab initio* calculations. *Int. Rev. Phys. Chem.* **2007**, *26*, 193−222.

(8) Misquitta, A. J.; Stone, A. J. Distributed polarizabilities obtained using a constrained density-fitting algorithm. *J. Chem. Phys.* **2006**, *124*, 024111.

(9) Williams, G. J.; Stone, A. J. Distributed dispersion: a new approach. *J. Chem. Phys.* **2003**, *119*, 4620−4628.

(10) Hulme, A. T.; Johnston, A.; Florence, A. J.; Fernandes, P.; Shankland, K.; Bedford, C. T.; Welch, G. W. A.; Sadiq, G.; Haynes, D. A.; Motherwell, W. D. S.; Tocher, D. A.; Price, S. L. The search for a predicted hydrogen bonded motif − a multidisciplinary investigation into the polymorphism of 3-azabicyclo[3.3.1]nonane-2,4-dione. *J. Am. Chem. Soc.* **2007**, *129*, 3649−3657.

(11) 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.

(12) Jeziorski, B.; Szalewicz, K. Symmetry-adapted perturbation theory. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley: 2002; Vol. 8, pp 37−83.

(13) Patkowski, K.; Jeziorski, B.; Szalewicz, K. Symmetry-adapted perturbation theory with regularized coulomb potential. *J. Mol. Struct.* (*THEOCHEM*) **2001**, *547*, 293−307.

(14) Patkowski, K.; Jeziorski, B.; Szalewicz, K. Unified treatment of chemical and van der Waals forces via symmetry-adapted perturbation expansion. *J. Chem. Phys.* **2004**, *120*, 6849−6862.

(15) Misquitta, A. J.; Stone, A. J. *CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form*; University of Cambridge: 2006. Inquiries to A. J. Misquitta, am592@cam.ac.uk.

(16) Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E.; Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. *Orient: a program for studying interactions between molecules, version 4.6*; University of Cambridge: 2006. Inquiries to A. J. Stone, ajs1@cam.ac.uk.

(17) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(18) Sadlej, A. J. Medium-size polarized basis sets for high-level correlated calculations of molecular electric properties. *Collect. Czech Chem. Commun.* **1988**, *53*, 1995−2016.

(19) Sadlej, A. J. Medium-sized polarized basis sets for high-level correlated calculations of molecular electric properties. II. Second-row atoms Si−Cl. *Theor. Chim. Acta* **1991**, *79*, 123−140.

(20) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. On the effectiveness of monomer-centered, dimer-centered, and bond-centered basis functions in calculations of intermolecular interaction energies. *J. Chem. Phys.* **1995**, *103*, 7374−7391.

(21) Patkowski, K.; Szalewicz, K.; Jeziorski, B. Third-order interactions in symmetry-adapted perturbation theory. *J. Chem. Phys.* **2006**, *125*, 154107.

(22) Misquitta, A. J.; Stone, A. J. Accurate dispersion energies for organic molecules. **2007**, manuscript in preparation.

(23) Le Sueur, C. R.; Stone, A. J. Localization methods for distributed polarizabilities. *Mol. Phys.* **1994**, *83*, 293−308.

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

(25) Hodges, M. P.; Stone, A. J. A new representation of the dispersion interaction. *Mol. Phys.* **2000**, *98*, 275−286.

(26) Tang, K. T.; Toennies, J. P. An improved simple model for the Van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* **1984**, *80*, 3726−3741.

(27) Levy, M.; Perdew, J. P.; Sahni, V. Exact differential equation for the density and ionization energy of a many-particle system. *Phys. Rev. A* **1984**, *30*, 2745−2748.

(28) Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. Is the induction energy important for modelling organic crystals. *J. Chem. Theory Comput.* **2007**, submitted for publication.

(29) Kim, Y. S.; Kim, S. K.; Lee, W. D. Dependence of the closed-shell repulsive interaction on the overlap of the electron densities. *Chem. Phys. Lett.* **1981**, *80*, 574−575.

(30) Burcl, R.; Chalasinski, G.; Bukowski, R.; Szczesniak, M. M. On the role of bond functions in interaction energy calculations: Ar···HCl, Ar···H_2O, $(HF)_2$. *J. Chem. Phys.* **1995**, *103*, 1498−1507.

(31) Langlet, J.; Caillet, J.; Bergès, J.; Reinhardt, P. Comparison of two ways to decompose intermolecular interactions for hydrogen-bonded dimer systems. *J. Chem. Phys.* **2003**, *118*, 6157−6166.

(32) Millot, C.; Stone, A. J. Towards an accurate intermolecular potential for water. *Mol. Phys.* **1992**, *77*, 439−462.

(33) Stone, A. J.; Alderton, M. Distributed multipole analysis—methods and applications. *Mol. Phys.* **1985**, *56*, 1047−1064.

(34) Stone, A. J. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.

(35) Adamo, C.; Cossi, M.; Scalmani, G.; Barone, V. Accurate static polarizabilities by density functional theory: assessment of the PBE0 model. *Chem. Phys. Lett.* **1999**, *307*, 265−271.

(36) Tozer, D. J.; Handy, N. C. Improving virtual Kohn−Sham orbitals and eigenvalues: Application to excitation energies and static polarizabilities. *J. Chem. Phys.* **1998**, *109*, 10180−10189.

(37) Tozer, D. J. The asymptotic exchange potential in Kohn−Sham theory. *J. Chem. Phys.* **2000**, *112*, 3507−3515.

(38) Bukowski, R.; Podeszwa, R.; Szalewicz, K. Efficient generation of the coupled Kohn−Sham dynamic sysceptibility functions and dispersion energy with density fitting. *Chem. Phys. Lett.* **2005**, *414*, 111−116.

(39) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Density-fitting method in symmetry-adapted perturbation theory based on Kohn−Sham description of monomers. *J. Chem. Theory Comput.* **2006**, *2*, 400−412.

(40) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143−152.

(41) Weigend, F.; Kohn, A.; Hättig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(42) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Kucharski, S.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Rybak, S.; Szalewicz, K.; Williams, H.; Wormer, P. *SAPT2002: an ab initio program for many-body symmetry-adapted perturbation theory calculations of intermolecular interaction energies*; University of Delaware and University of Warsaw: 2002.

(43) *DALTON, a molecular electronic structure program, release 2.0*; 2005. See: http://www.kjemi.uio.no/software/dalton/dalton.html (accessed 21 June 2007).

CT700105F

JCTC Journal of Chemical Theory and Computation

Evaluation of the Electrostatically Embedded Many-Body Expansion and the Electrostatically Embedded Many-Body Expansion of the Correlation Energy by Application to Low-Lying Water Hexamers

Erin E. Dahlke, Hannah R. Leverentz, and Donald G. Truhlar*

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

Received July 24, 2007

Abstract: We have applied a many-body (MB) expansion, the electrostatically embedded many-body (EE-MB) approximation, and the electrostatically embedded many-body expansion of the correlation energy (EE-MB-CE), each at the two-body (MB = PA, where PA denotes pairwise additive) and three-body (MB = 3B) levels, to calculate total energies for a series of low-lying water hexamers using eight correlated levels of theory including second-order and fourth-order Møller−Plesset perturbation theory (MP2 and MP4) and coupled cluster theory with single, double, and quasipertubative triple excitations (CCSD(T)). Comparison of the expansion methods to energies obtained from full (i.e., unexpanded) calculations shows that the EE-3B-CE method is able to reproduce the full cluster energies to within 0.03 kcal/mol, on average. We have also found that the deviations of the results predicted by the expansion methods from those obtained with full calculations are nearly independent of the correlated level of theory used; this observation will allow validation of the many-body methods on large clusters at less expensive levels of theory (such as MP2) to be extrapolated to the CCSD(T) level of theory. Furthermore, we have been able to rationalize the accuracies of the MB, EE-MB, and EE-MB-CE methods for the six hexamers in terms of the specific many-body effects present in each cluster.

1. Introduction

The ability to calculate accurate energies for large systems has long been a goal of the quantum chemical community. Hartree−Fock theory, which neglects electron correlation, is inadequate but is used to generate orbitals for methods such as second-order Møller−Plesset perturbation theory, MP2,[1] coupled cluster theory with single and double excitations, CCSD,[2,3] or CCSD with quasiperturbative connected triple excitations, CCSD(T),[4] which do include electron correlation and are able to accurately predict energies, geometries, and frequencies for small to moderately sized chemical systems. However, these post-Hartree−Fock methods have thus far proven to be too expensive, in their original implementations, to be used for systems containing tens to hundreds of atoms. As a result, there has been considerable research aimed at trying to develop highly efficient alogorithms,[5−9] including parallelization schemes,[10−18] to make large systems tractable. However, because of the steep scaling of computational effort with respect to system size (CCSD(T), CCSD, and MP2 scale as N^7, N^6, and N^5, respectively, where N is the number of atoms[19]), it is impractical to utilize even these more efficient implementations for systems containing hundreds to thousands of atoms.

A promising area of research has focused on developing variants of these methods that use localized molecular orbitals[15,20−23] or fragmentation.[13,24−40] One can also consider including a subset of interactions, for example Coulomb interactions, to high order or in full, with other interactions, e.g., those due to electron correlation energy considered only to a lower order, e.g., only pairwise.[41−43] In past work[44,45] we have developed the electrostatically embedded many-body method (EE-MB) and the electrostatically embedded

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

10.1021/ct700183y CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/08/2007

many-body expansion of the correlation energy (EE-MB-CE) methods and have applied both methods to the study of water clusters, ranging in size from 5 to 20 water molecules, at the MP2 level of theory. MP2 was chosen because it is the least expensive of the post-Hartree−Fock methods, allowing for direct comparison of the EE-MB and EE-MB-CE methods to the MP2 energies of the full cluster. In this way we were able to examine the performance of the expansion methods with respect to increasing system size, which would not have been possible with other more expensive post-Hartree−Fock methods like CCSD(T).

Despite the advantages of using MP2 for our initial studies, there is a concern that other levels of electronic structure theory could show different behavior. Because of the N^6 or N^7 scaling of the post-MP2 methods with respect to system size, if we wish to compare directly to full calculations (i.e., conventional calculations without a many-body expansion) at a post-MP2 level with a reasonably large basis set (polarized valence triple-ζ or higher) we are limited to relatively small systems (on the order of five heavy atoms). Recent work by Olson et al.[18] has provided CCSD(T), CCSD, and MP2 energies for a series of five water hexamers using both the aug-cc-pVTZ[46,47] and s-cc-pVTZ basis sets (s-cc-pVTZ denotes semidiffuse cc-pVTZ, and it uses the aug-cc-pVTZ basis set on oxygen and the cc-pVTZ[48] basis set on hydrogen). These results constitute a set of highly accurate energies against which to test our methods.

Water clusters in general exhibit large many-body effects,[49] and it is well-known that different structural motifs can lead to different many-body effects;[50,51] therefore, these clusters are an excellent choice for examining the behavior of many-body methods. Along this line, Pedulla and Jordan[51] have examined the many-body effects of three isomers of the water hexamer (cage, prism, and ring). Since there are a number of isomers of the water hexamer that all lie within a few kilocalories per mole of each other,[18,52−58] this system serves as a good test of the predictive capabilities of electronic structure methods for water.

For any level of theory (e.g., MP2 or CCSD(T) with a given basis) we can either perform full (i.e., conventional) calculations of the potential energy, V, or many-body expansions, with the latter defined by a truncated version of

$$V = V_1 + V_2 + V_3 + \cdots \tag{1}$$

where V_n is the n-body term. Truncating at V_2 is called the pairwise additive (PA) approximation, and truncating at V_3 is called the three-body (3B) approximation. For $(H_2O)_N$, V_2 involves calculating $(N(N-1)/2)$ dimer calculations, and V_3 involves $(N(N-1)(N-2)/3!)$ trimer calculations. If the n-mer calculations are performed in vacuum, we have a conventional many-body method (PA or 3B), and if they are performed in a field of point charges at the nuclear positions of the $N - n$ missing monomers, we have the electrostatically embedded many-body method (EE-PA or EE-3B). If we perform a full (V_N) calculation at the Hartree−Fock (HF) level and expand the correlation energy ($V -$ V$_{\mathrm{HF}}$), we have the many-body expansion of the correlation energy method (PA-CE or 3B-CE without point charges and EE-PA-CE or EE-3B-CE with them). Further details of the many-body methods have been discussed in previous work[44,45] and are not discussed here.

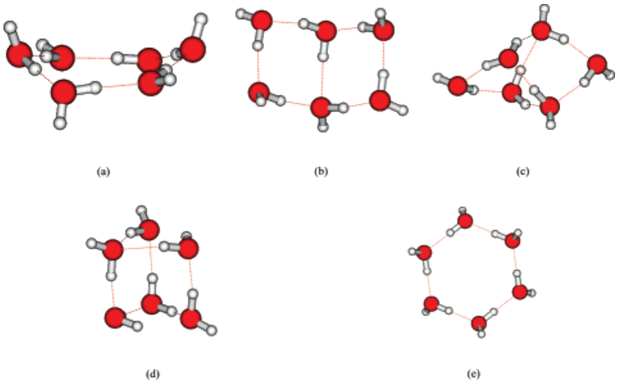


Figure 1. MP2/DH(d,p) optimized hexamers of Day et al. for the (a) boat, (b) book, (c) cage, (d) prism, and (e) ring structures.

2. Computational Details

The hexamers used in this work are defined by the MP2/DH(d,p) geometries of Day et al.[55] for the boat, book, cage, ring (denoted as cyclic in that work), and prism isomers (see Figure 1). All single-point calculations in this work use the s-cc-pVTZ basis set (see section 1 for the definition of s-cc-pVTZ). For each hexamer a total of nine levels of electronic structure theory were considered: Hartree−Fock, MP2, MP3, MP4D, MP4DQ, MP4SDQ, MP4, CCSD, and CCSD(T), where MP3, MP4D, MP4DQ, MP4SDQ, and MP4 denote various high-order perturbation theory approximations.[59−62] The CCSD(T), CCSD, and MP2 single-point energies for the hexamers were taken from the work of Olson et al.[18] In addition, MP4 calculations were run with the *Gaussian 03*[63] software program to determine the MP3, MP4D, MP4DQ, MP4SDQ, and MP4 single-point energies for each hexamer. PA, 3B, EE-PA, EE-3B, PA-CE, 3B-CE, EE-PA-CE, and EE-3B-CE calculations were carried for each hexamer, at each of the nine levels of theory, using the *Gaussian 03* software package.

For the sake of clarity a combination of many-body method and electronic structure theory will be denoted by the name of the many-body method with the level of electronic structure level in parentheses. For example, EE-PA-CE(MP2) will denote an EE-PA-CE calculation carried out at the MP2 level of theory.

In the EE-MB and EE-MB-CE methods, charges of −0.778 and 0.389 were used for the oxygen and hydrogen atoms, respectively, as in refs 44 and 45.

3. Results and Discussion

3.1. Full Calculations. While the main goal of this paper is analysis of the ability of the many-body methods to accurately reproduce full energies obtained at the same level of electronic structure theory (e.g., comparing a EE-PA-CE-(MP2) calculation to a full MP2 calculation on the same hexamer), it is useful to first examine the results of the full calculations. Table 1 shows the relative energy differences (relative to the prism structure) predicted by the full calculations for each level of theory. All eight correlated

Table 1. Relative[a] Energies (kcal/mol) Predicted by Full Calculations Using the s-cc-pVTZ Basis Set[b]

	boat	book	cage	prism	ring
HF	−1.79	−1.05	−0.27	0.00	−2.52
MP2	2.41	0.67	0.02	0.00	1.22
MP3	2.25	0.74	0.13	0.00	1.15
MP4D	2.33	0.76	0.14	0.00	1.21
MP4DQ	1.92	0.60	0.10	0.00	0.84
MP4SDQ	2.20	0.71	0.12	0.00	1.09
MP4	2.99	1.02	0.17	0.00	1.80
CCSD	2.23	0.74	0.14	0.00	1.12
CCSD(T)	2.99	1.06	0.20	0.00	1.81

[a] All energies are relative to the prism isomer [b] The s-cc-pVTZ basis set denotes the aug-cc-pVTZ basis set on oxygen and the cc-pVTZ basis set on hydrogen.

Table 2. Average Deviations between Various Full Calculated Energies and Full CCSD(T)[a]

	MUE	RMSE
HF	2.68	3.08
MP2	0.31	0.36
MP3	0.41	0.47
MP4D	0.37	0.43
MP4DQ	0.60	0.69
MP4SDQ	0.44	0.51
MP4	0.02	0.03
CCSD	0.43	0.49

[a] These results are averaged over the ten energy differences that can be obtained from the five structures in Table 1.

levels of theory predict a relative energy ordering of prism < cage < book < ring < boat. The energy differences separating the isomers are all less than 3 kcal/mol, with the smallest energy gap (cage minus prism) in the range between 0.02 and 0.20 kcal/mol.

An interesting result in Table 1 is how the perturbation theory and CCSD results compare to the CCSD(T) results. As was noted by Olson et al.[18] the MP2 results differ from the CCSD(T) results by 0.2−0.6 kcal/mol, with CCSD(T) predicting systematically larger energy gaps than MP2. Past work[57,64,65] comparing MP2 and CCSD(T) for small water clusters showed that the differences obtained are small (on the order of 0.05 kcal/mol for the water dimer and 0.1 kcal/mol for the trimer), and, as a result, MP2 has become the method of choice for many workers when studying water clusters.[51,56,65−68] The results in Table 1 indicate that the exceptionally good agreement of MP2 and CCSD(T) for small water clusters may begin to break down as larger clusters are considered, and thus caution should be used when applying MP2 to medium- to large-sized clusters. Another interesting result is how well MP4 is able to reproduce the CCSD(T) relative energies. Of the methods in Table 1, only MP4 and CCSD(T) include connected triple excitations. CCSD(T) includes not only the fourth-order connected triples of MP4 but also a fifth-order connected triple excitation operator involving singles amplitudes. The near equivalence of MP4SDQ and CCSD in Table 1 shows that disconnected triples (included in the latter but not the former) are unimportant for water clusters, and the near agreement of MP4 with CCSD(T) but not with MP4SDQ shows that for water clusters the connected triples are important, but it is adequate to include them at fourth order.

With five different structures there are a total of 10 energy differences that one can compute (for example, the energy difference between the boat and the book or between the cage and the ring). In order to better characterize how accurate the MPn ($n = 2-4$) and CCSD results are, as compared to CCSD(T), we have calculated the 10 energy differences at each of the 9 levels of theory and computed the mean unsigned and root mean squared errors relative to the CCSD(T) results. The results of this analysis are shown in Table 2. From this table it is clear that MP4 is very accurate for these systems, as compared to CCSD(T); it has a mean unsigned deviation of only 0.02 kcal/mol, and the largest difference between the MP4 and CCSD(T) for any of the 10 possible energy differences is 0.04 kcal/mol. This conclusion is in good agreement with previous work by Xantheas et al.[69] on the water dimer. While MP2 has the second-lowest mean unsigned error (0.31 kcal/mol), its performance is considerably worse than that of MP4. The remaining correlated methods have mean unsigned errors ranging from 0.37 to 0.60 kcal/mol. Hartree−Fock does particularly poorly, as is expected from the results in Table 1. While both MP4 and CCSD(T) formally scale as N^7, the use of CCSD(T) requires the completion of a CCSD calculation (which has an iterative N^6 step[4]) before the noniterative triples calculation. As a result, MP4 is less expensive than CCSD(T) which may allow MP4 to be used for benchmark calculations of water clusters that are too large for CCSD(T). It seems worthwhile to note that although CCSD(T) is well-known[4] to be more accurate than MP4 in general, where singles amplitudes may be large, this need not be the case for particular interactions of noncovalent interactions of closed-shell species.[70] Even though the present comparison involves larger clusters than those for which MPn results ($n = 2$, 3, 4) have previously (prior to ref 18) been compared to CCSD(T) results, and hence involves comparisons that are assumed to be more relevant to the bulk water case, there is no guarantee that MP4 is better than MP2 in general (the series is often divergent, especially with basis sets containing diffuse functions); and the good agreement of MP4 with CCSD(T) for these small water clusters is not guaranteed to hold for larger clusters. Therefore caution must be exercised in choosing either MP2 or MP4 as an alternative to CCSD(T) for water clusters. Further, systematic validation studies would be useful.

3.2. Pairwise Additive Methods. In order to assess the accuracy of the many-body methods we begin by evaluating the average error in the electronic energy for each pairwise additive method, at each level of electronic structure theory, when compared to the full calculation (i.e., evaluating the deviation between the EE-PA(MP2) and full MP2 energy for each of the five hexamers). Table 3 shows the average deviations between the pairwise additive methods (PA, PA-CE, EE-PA, EE-PA-CE) and the full calculations at each of the nine levels of theory. The first interesting observation is that the errors for each pairwise additive method (PA or EE-PA) are similar for all the correlated levels of theory as indicated by the standard deviations being much smaller than the average MUE. A comparison of the average mean

Table 3. Average Deviations[a] (kcal/mol) between Pairwise Additive Energies and Full Calculations at the Same Level of Theory

	PA			PA-CE			EE-PA			EE-PA-CE		
	MSD	MUD	RMSD	MSD	MUD	RMSD	MSD	MUD	RMSD	MSD	MUD	RMSD
HF	11.81	11.81	11.94	0.00	0.00	0.00	1.13	1.13	1.15	0.00	0.00	0.00
MP2	11.89	11.89	12.01	0.03	0.09	0.10	1.10	1.10	1.11	−0.04	0.04	0.04
MP3	11.75	11.75	11.89	0.01	0.11	0.13	0.96	0.96	1.01	−0.17	0.17	0.22
MP4D	11.75	11.75	11.89	0.01	0.10	0.13	0.96	0.96	1.00	−0.17	0.17	0.22
MP4DQ	11.70	11.70	11.83	−0.03	0.12	0.14	0.95	0.95	0.98	−0.19	0.19	0.21
MP4SDQ	11.72	11.72	11.86	−0.02	0.10	0.12	0.98	0.98	1.02	−0.15	0.15	0.18
MP4	11.80	11.80	11.93	0.03	0.08	0.09	1.02	1.02	1.06	−0.11	0.11	0.15
CCSD	11.74	11.74	11.88	−0.01	0.09	0.11	0.99	0.99	1.02	−0.14	0.14	0.18
CCSD(T)	11.82	11.82	11.96	0.05	0.09	0.10	1.03	1.03	1.07	−0.11	0.12	0.16
av MUE[b]		11.77			0.10			1.00			0.14	
SD[c]		0.06			0.01			0.05			0.05	

[a] MSD, MUD, and RMSD denote mean signed, mean unsigned, and root-mean-squared deviations, respectively, in *V* as compared to the full calculations. Thus a positive MSE corresponds to underestimating the strength of binding, and a negative MSE corresponds to overestimating the strength of binding. [b] Average of the MUE for the correlated methods (the rows from MP2 to CCSD(T)). [c] Standard deviation of MUE for the correlated methods (the rows from MP2 to CCSD(T)).

Table 4. Average Deviations[a] (kcal/mol) between Three-Body (3B) Methods and Full Calculations at the Same Level of Theory

	3B			3B-CE			EE-3B			EE-3B-CE		
	MSD	MUD	RMSD	MSD	MUD	RMSD	MSD	MUD	RMSD	MSD	MUD	RMSD
HF	1.08	1.08	1.22	0.00	0.00	0.00	0.08	0.09	0.11	0.00	0.00	0.00
MP2	1.25	1.25	1.42	0.17	0.17	0.20	0.09	0.12	0.15	0.01	0.03	0.04
MP3	1.21	1.21	1.37	0.12	0.12	0.15	0.08	0.11	0.14	0.00	0.03	0.03
MP4D	1.23	1.23	1.39	0.15	0.15	0.17	0.09	0.11	0.14	0.01	0.03	0.03
MP4DQ	1.21	1.21	1.37	0.13	0.13	0.15	0.08	0.11	0.13	0.00	0.03	0.03
MP4SDQ	1.23	1.23	1.39	0.15	0.15	0.17	0.08	0.11	0.14	0.00	0.03	0.04
MP4	1.28	1.28	1.45	0.20	0.20	0.23	0.09	0.13	0.16	0.01	0.04	0.05
CCSD	1.22	1.22	1.39	0.14	0.14	0.17	0.08	0.11	0.14	0.00	0.03	0.04
CCSD(T)	1.27	1.27	1.44	0.19	0.19	0.22	0.09	0.13	0.16	0.01	0.05	0.05
av MUE[b]		1.24			0.16			0.12			0.03	
SD[c]		0.03			0.03			0.01			0.01	

[a] MSD, MUD, and RMSD denote mean signed, mean unsigned, and root-mean-squared deviations, respectively, as compared to full calculations. See footnote *a* of Table 3 for an explanation of the signs. [b] Average of the MUE for the correlated methods (the rows MP2 to CCSD(T)). [c] Standard deviation of the correlated methods (the rows MP2 to CCSD(T)).

unsigned errors shows that the PA method has a large error of 11.78 kcal/mol. The average binding energy of the six structures at the CCSD(T) level is 46.72 kcal/mol (taken from ref 18), so an average error of 11.78 kcal/mol corresponds to a percent error of approximately 25% (the next largest percent error is 2.2% for the EE-PA method). The EE-PA method shows an order-of-magnitude improvement over the PA method, and the PA-CE and EE-PA-CE methods have errors that are 2 orders of magnitude better.

Furthermore, one can see that for the PA and EE-PA methods the mean unsigned error for the Hartree−Fock level of theory is as large or larger than the errors for the correlated methods (by definition the Hartree−Fock errors for the PA-CE and EE-PA-CE methods are zero because they include a full Hartree−Fock calculation). The fact that the errors in the PA-CE and EE-PA-CE methods are much smaller than errors in the PA and EE-PA methods indicates that the largest breakdown in the many-body and electrostatically embedded many-body expansions are in the Hartree−Fock energy and that even a pairwise treatment of the correlation energy is sufficient to reproduce the correlation energy at a given level to within 0.2 kcal/mol. Given that even the CCSD(T) level of theory is believed to be accurate only to within ∼1 kcal/mol the difference between the pairwise methods at a given level of electronic structure theory (with the exception of the standard pairwise additive approximation) and the conventional calculation (i.e., a calculation on the full cluster at the same level of theory) is expected to be of the same magnitude or smaller than the difference between the conventional calculation and the result from a full configuration interaction calculation.

3.3. Three-Body Methods. We continue our analysis by comparing the electronic energies predicted by the three-body method, at each level of electronic structure theory, to the full calculation for each of the five hexamers. Table 4 shows the average errors between the three-body methods (3B, 3B-CE, EE-3B, EE-3B-CE) and the full calculations at each level of theory. As expected,[45] we see that the 3B calculations are approximately an order of magnitude better than the PA results (compare Table 4 to Table 3). Also, we see that including the full Hartree−Fock energy reduces the errors by approximately an order of magnitude as one goes from the 3B to the 3B-CE method and from the EE-3B to the EE-3B-CE method. We again see that all of the correlated methods have very similar errors for each of the many-body methods and that the standard deviations are even lower for

Table 5. Signed Errors (kcal/mol) for the Many-Body Methods for the Five Hexamer Structures Relative to the Full CCSD(T) Calculation

	boat	book	cage	prism	ring	MUE[b]
PA[a]	13.67	11.71	9.93	9.74	14.06	2.47
3B	1.99	1.14	0.52	0.59	2.11	0.92
PA-CE	−0.01	0.13	−0.10	0.13	0.09	0.14
3B-CE	0.30	0.19	0.06	0.06	0.32	0.15
EE-PA	1.25	1.06	0.78	0.65	1.40	0.39
EE-3B	0.21	0.08	−0.01	−0.09	0.25	0.18
EE-PA-CE	−0.01	−0.05	−0.23	−0.28	0.04	0.17
EE-3B-CE	0.05	0.02	−0.02	−0.06	0.07	0.07

[a] In this table, PA denotes PA(CCSD(T)), EE-PA denotes EE-PA(CCSD(T)), etc. [b] MUE denotes mean unsigned error of the ten comparisons of the relative energies for two structures; see section 3 of the text for an explanation.

the three-body methods than at the pairwise additive level. The 3B-CE errors are smaller than the 3B ones, and the EE-3B-CE errors are smaller than the EE-3B errors, indicating again that the dominant errors associated with the 3B and EE-3B methods are due to the Hartree–Fock energy and not the correlation energy. As seen with the pairwise additive methods, the errors associated with the three-body methods (with the possible exception of the conventional three-body method, 3B) are smaller than the intrinsic errors of the levels of electronic structure theory that are tested.

3.4. Analysis of the Many-Body Methods. Based on the results in Tables 3 and 4, the many-body methods can be ranked in order of decreasing mean unsigned error as PA ≫ 3B > EE-PA ≫ 3B-CE > EE-PA-CE > EE-3B > PA-CE ≫ EE-3B-CE; however, Tables 3 and 4 deal only with average over errors. In order to truly understand the sources of the errors in the many-body methods, it is useful to also look at the individual error for each hexamer. Because all the correlated electronic structure methods in Table 2 show very similar results for each many-body method, only one level of electronic structure theory need be discussed in this respect. We will discuss the analysis of the CCSD(T) case, for which the key results are given in Table 5. The first set of results shown in Table 5 is the error in absolute energy for each hexamer, calculated with each many-body method, compared to the full CCSD(T) calculation (for example, the error between the energy of the cage isomer calculated at the PA(CCSD(T)) level of theory and the energy of the cage isomer calculated using a full CCSD(T) calculation). As mentioned in section 3.1, with 5 hexamers there are 10 total energy differences that one can compute (for example, the energy difference between the boat and the book or between the cage and the ring). We have calculated these 10 energy differences for each many-body method listed in Table 5 and compared them to the results from the full CCSD(T) calculations; the resulting mean unsigned error can be found in the last column of Table 5. The pupose of this analysis is not to assess the accuracy of the many-body methods (which has been done in the two previous sections) but to see if we can gain any insight into the performance of the many-body methods with respect to the many-body terms present in the structures.

Table 5 shows that the mean unsigned errors for the PA-CE, 3B-CE, EE-3B, and EE-PA-CE methods all lie within 0.04 kcal/mol of each other—as was stated in our previous discussions of Tables 2 and 3—which make them all appropriate for use on systems that require high accuracy; however, if the errors for individual structures are compared, the four methods behave quite differently. For example, the EE-PA-CE method has smaller errors for the boat, book, and ring structure than for for the cage and prism, whereas the EE-3B method has smaller errors for the book, cage, and prism than for the boat and ring structures.

Because each method in Table 5 uses a different approximation to calculate the many-body effects in these clusters (i.e., neglecting some terms or including them in an average way via the point charges) it is reasonable to assume that their performance for the hexamers is directly related to the many-body effects present in the hexamers. Therefore, in order to better understand any systematic shortcomings of each many-body method we must first have a good understanding of the many-body effects in each structure.

3.4.1. General Discussion of Many-Body Effects. Before we begin this analysis we will take a moment to clarify a few terms necessary in our discussion. First of all, when we refer to smaller errors, we mean smaller absolute values of errors. Second, within a many-body expansion the total energy of the system is written as a sum of n-body terms denoted by V_n (see eq 1 in the Introduction) in which the one-body (V_1), two-body (V_2), and three-body (V_3) terms can be written as

$$V_1 = \sum_i E_i \tag{2}$$

$$V_2 = \sum_{i<j} (E_{ij} - E_i - E_j) \tag{3}$$

$$V_3 = \sum_{i<j<k} [(E_{ijk} - E_i - E_j - E_k) - (E_{ij} - E_i - E_j) - (E_{ik} - E_i - E_k) - (E_{jk} - E_j - E_k)] \tag{4}$$

respectively, and so on for higher-order terms, and where E_i, E_{ij}, and E_{ijk}, ..., are the energies of the monomers, dimers, trimers, and so forth, in the system.

Because the electronic energies of all the monomers in the system are negative, each term of the series in eq 2 is negative, and, therefore, V_1 must be negative. For the series in eqs 3 and 4, however, each term may be positive or negative. For example, if the energy of dimer E_{ij} is higher in energy than the sum of its constituent monomer energies (i.e., an unfavorable interaction) the corresponding term in the series will be positive.

Throughout this analysis if we are talking about the series in eqs 2, 3, or 4 we will refer to them as the one-body, two-body, or three-body terms. If we are talking about the individual terms making up these series we will refer to them as an individual one-body, individual two-body, or individual three-body terms. We will also use the phrase “beyond-three-body terms” to denote the sum of the four-, five-, and six-body terms.

Table 6. Two-, Three-, and Beyond-Three-Body Terms (kcal/mol) at the CCSD(T) Level of Theory

	V_2	V_3	beyond V_3[a]
boat	−31.99	−11.68	−1.99
book	−36.07	−10.57	−1.14
cage	−38.62	−9.41	−0.52
prism	−39.06	−9.16	−0.59
ring	−32.82	−11.95	−2.11

[a] Beyond V_3 denotes the sum of four-, five-, and six-body terms.

Table 7. Contribution of the Correlation Energy to the Two-, Three-, and Beyond-Three-Body Terms (kcal/mol) at the CCSD(T) Level of Theory

	V_2	V_3	beyond V_3[a]
boat	−11.87	0.25	−0.30
book	−14.24	0.06	−0.19
cage	−16.15	0.16	−0.06
prism	−16.43	0.19	−0.06
ring	−12.08	0.24	−0.32

[a] Beyond V_3 denotes the sum of four-, five-, and six-body terms.

3.4.2. Many-Body Effects in Water Hexamers. To analyze the many-body effects of each hexamer we will build on the insights of Pedulla and Jordan[51] who have carried out a many-body analysis using the MP2 level of theory with the aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ basis sets for the cage, ring, and prism isomers (optimized at the MP2/6-31+G(2d,p) level of theory). The relevant conclusions of ref 51 are as follows: (i) the cage and prism have larger two-body terms than the ring due to the presence of more hydrogen bonds, (ii) the ring isomer has larger three- and four-body terms because all of the individual two- and three-body terms have the same sign, (iii) five- and six-body terms are ≤0.05 kcal/mol for the cage and prism structure but are as large as 0.20 kcal/mol for the ring, and (iv) the effects of electron correlation are relatively unimportant for many-body terms beyond third order.

Table 6 shows the two-, three-, and beyond-three-body terms, at the CCSD(T)/s-cc-pVTZ level of theory, for each hexamer considered in this work. Following the work of Pedulla and Jordan we have also computed the contribution of the correlation energy to the two-, three-, and beyond-three-body terms; these results are shown in Table 7. Tables 6 and 7 show that the results obtained for our clusters at the CCSD(T)/s-cc-pVTZ level of theory are consistent with the work of Pedulla and Jordan at the MP2/aug-cc-pVTZ level of theory. We see that the three-body and beyond-three-body terms are approximately four times larger for the ring and the boat structure than for the cage and the prism and that the book structure is intermediate between these two groups. We also see that the contribution of the correlation energy to the three-body and beyond-three-body terms is approximately 2 orders of magnitude smaller than its contribution to the two-body terms, and finally we see that the magnitudes of the three-body and beyond-three-body terms are very similar.

3.4.3. Analysis of the Nonelectrostatically Embedded Methods. Table 6 shows that the terms beyond the two-body terms are smallest for the prism and cage structure and largest for the boat and ring structure. Therefore, the PA method (which neglects all three-body and higher terms) will perform the best for the prism and the worst for the ring, which is confirmed by Table 5. At the 3B level of theory the errors are significantly reduced compared to the PA method (because only four-body and higher terms are neglected), but the results are still best for structures with small four-, five-, and six-body effects (i.e., prism and cage) and worst for structures with larger many-body effects (i.e., ring); this agrees with the results in Table 5.

At the PA-CE level, $V_1 - V_6$ are accounted for at the Hartree−Fock level, but correlation effects are considered only for the one- and two-body terms. The work of Pedulla and Jordan and the results of Table 7 show that inclusion of correlation energy has only a relatively small effect on beyond-three-body terms. As a result Table 5 shows that the PA-CE errors are much lower (nearly 2 orders of magnitude lower) than the PA errors (due to inclusion of $V_3 - V_6$ at the Hartree−Fock level). Table 5 also shows that the PA-CE method performs better than the 3B method (due to inclusion of $V_4 - V_6$ at the Hartree−Fock level). Finally, for the 3B-CE method one would expect improved performance over the three previously discussed methods, because only the four-, five-, and six-body correlation terms are neglected; however, the errors for the boat and ring are substantially larger at the 3B-CE level than at the PA-CE level; however, this can be explained by examining the contributions of correlation energy to the three-body and beyond-three-body terms. Table 7 shows that if only correlation effects are considered, the magnitude of the three-body and beyond-three-body terms are similar, but that they have different signs. Because these terms are nearly equal and opposite when both are neglected (i.e., in the PA-CE method) the errors cancel each other, and the overall error is lower than may have been expected; however, when only the latter is neglected (i.e., in the 3B-CE method) there is no such cancellation and the errors increase, particularly for structures like the ring and boat.

3.4.4. Analysis of the Electrostatically Embedded Methods. In the EE-PA approximation all two-body terms are taken into account explicitly, and the beyond-pairwise terms are accounted for in an average way by the presence of the point charges. As a result, Table 5 shows that the overall errors are substantially (approximately 1 order of magnitude) smaller than for the PA method. As a further consequence, the EE-PA method performs best for the prism and cage (which have smaller three-body and beyond-three-body terms). Table 5 also shows that the EE-PA method has lower errors than the 3B method for the ring, boat, and book structures and slightly higher errors for the cage and prism structure. These results are most likely due to not explicitly accounting for the three-body correlation terms discussed at the end of the previous section. The EE-3B method explicitly includes the three-body correlation terms, and, as a result, the errors are reduced by nearly an order of magnitude compared to the EE-PA method.

Table 5 also shows that EE-PA-CE method, by including the full Hartree−Fock energy, has smaller errors than the EE-PA method; this result is expected based on the non-EE results. The non-EE results also suggest that the largest errors

Table 8. Timings[a] for MP4 and Many-Body Methods at the CCSD(T) Level of Theory, Relative to MP2 with the Same s-cc-pVTZ Basis Set[b]

	timing
MP4	178
PA or EE-PA	5.0
3B or EE-3B	110
PA-CE or EE-PA-CE	5.4
3B-CE or EE-3B-CE	111

[a] All calculations use the s-cc-pVTZ basis set. [b] In this table, PA denotes PA(CCSD(T)), EE-PA denotes EE-PA(CCSD(T)), etc.

should be for the cage and prism (due to not explicitly including the three-body correlation energy), which is consistent with the results in Table 5. Table 5 also shows that the EE-PA-CE errors for the ring, boat, and book structures are lower than the EE-3B errors, while the opposite is true for the cage and prism. This result can be rationalized by considering the largest error in each method, in particular failing to include the full Hartree−Fock energy in the EE-3B method and not explicitly accounting for the three-body correlation terms in the EE-PA-CE method. The EE-MB-(HF) errors for the boat, book, cage, prism, and ring are 0.16, 0.06, 0.01, −0.02, and 0.18 kcal/mol, whereas the errors associated with not explicitly accounting for the electrostatically embedded three-body correlation energy are −0.06, −0.07, −0.21, −0.21, and −0.03 kcal/mol. The dominant errors for the two methods have quite different effects on the different structures. Based on this observation one might have predicted that if an EE-3B expansion of the Hartree−Fock energy were used with an EE-PA expansion of the correlation energy that the errors for the book, boat, cage, prism, and ring would be 0.10, −0.01, −0.19, −0.24, and 0.15 kcal/mol, respectively (assuming the errors are purely additive); the actual errors obtained are 0.15, 0.01, −0.21, −0.30, and 0.22 kcal/mol.

In the EE-3B-CE method, $V_1 - V_6$ are accounted for explicitly in the Hartree−Fock energy, the contribution to the two- and three-body terms from the correlation energy is explicitly taken into account, and the contribution of correlation energy to the higher-order terms is included in an average way. As a result, the EE-3B-CE method has the lowest errors; in particular, Table 5 shows that the error at the EE-3B-CE level has a magnitude of 0.07 kcal/mol or less for all five of the hexamers.

3.5. Timings. In order to evaluate the usefulness of the EE-MB and EE-MB-CE expansions, we must consider not only their accuracies but also their costs relative to competitive, or potentially competitive, methods. Therefore, we have computed the average times needed, on a single processor, to calculate the hexamer energies at the MP4/s-cc-pVTZ levels of theory and also for the eight many-body methods at the CCSD(T)/s-cc-pVTZ level of theory and have expressed these timings relative to the time needed to calculate the same energies at the MP2/s-cc-pVTZ level of theory with the same computer program and on the same computer (note that even ratios of timings depend on the computer program and computer). These timings are given in Table 8. First, the table shows that all eight many-body methods at the CCSD(T) level of theory are less expensive than an MP4 calculation on the full system. As the system size increases the many-body methods will become increasingly cost-effective relative to full MP4 calculations. Second, inclusion of the point charges does not change the amount of time needed to carry out the many-body expansion. Third, inclusion of the full Hartree−Fock energy causes a negligible increase in cost for these small systems.

Perhaps most importantly is that the pairwise methods are only five times more expensive than an MP2 calculation. This is important because both the PA-CE(CCSD(T)) and EE-PA-CE(CCSD(T)) methods perform better than MP2 when compared to the full CCSD(T) calculations. While the three-body methods are approximately 100 times more expensive than full MP2 calculations, all of the monomer, dimer, and trimer calculations are independent of each other which allows them each to be run on a different processor. As a result, the many-body methods are all highly parallelizable and so for moderately sized systems (on the order of 10 monomers) can be run in under a day, even at the CCSD-(T) level of theory.

4. Conclusions

Many-body (MB), electrostatically embedded many-body (EE-MB), and electrostatically embedded many-body expansion of the correlation energy (EE-MB-CE) calculations were carried out on five low-lying water hexamers and compared to full calculations at eight correlated levels of electronic structure theory ranging from MP2 theory to CCSD(T). We found that the average absolute errors associated with the many-body methods are consistent over the correlated levels of theory tested. Furthermore, when the errors obtained with each many-body method for each structure are compared they are also consistent across all levels of theory.

The mean unsigned errors in the relative energies of the structures are 0.17 and 0.07 kcal/mol, respectively, for EE-PA-CE and EE-3B-CE calculations, as compared to mean unsigned errors of 2.47 and 0.92 kcal/mol for conventional PA and 3B calculations, although the EE improvement adds negligibly to the computational cost, and the CE improvement requires only adding a Hartree−Fock calculation of the full system (which, for small- or moderate-sized systems is negligible or small, respectively). Finally, if one compares the accuracy of the many-body methods for reproducing the CCSD(T) relative (between structures) energy differences to the accuracy of full MP2 calculations (where "full" denotes without a many-body expansion, and where we note that full MP2 is commonly used in the literature for water clusters), we find that carrying out EE-MB, MB-CE, and EE-MB-CE calculations at the CCSD(T) level gives far better results, despite the method being competitive in speed if the many-body methods are run in parallel. We have also found that MP2 appears to be somewhat anomalous in that it is the only method that has a lower mean unsigned error for the EE-PA-CE method than for the EE-3B-CE method, which is probably just an accident.

In addition, we have been able to rationalize the performance of the EE-MB and EE-MB-CE methods on the various isomers in terms of the many-body effects of the clusters themselves. This will allow us to use the most cost-effective method possible for future studies and can provide insight into the performance of these methods on other systems.

Acknowledgment. The authors are grateful to Ryan Olson for an advance copy of ref 18. This work was supported in part by the National Science Foundation under grant nos. ITR-0428774 and CHE-0704974.

References

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

(2) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.

(3) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

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

(5) Scuseria, G. E.; Lee, T. J.; Schaefer, H. F., III *Chem. Phys. Lett.* **1986**, *130*, 236.

(6) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503−506.

(7) Scuseria, G. E.; Scheiner, A. C.; Lee, T. J.; Rice, J. E.; Schaefer, H. F., III *J. Chem. Phys.* **1987**, *86*, 2881.

(8) Lee. T. J.; Rice, J. E. *Chem. Phys. Lett.* **1988**, *150*, 406.

(9) Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *94*, 4334.

(10) Rendell, A. P.; Lee, T. J.; Komornicki, A. *Chem. Phys. Lett.* **1991**, *178*, 462.

(11) Rendell, A. P.; Lee, T. J.; Lindh, R. *Chem. Phys. Lett.* **1992**, *194*, 845.

(12) Rendell, A. P.; Guest, M. F.; Kendall, R. A. *J. Comput. Chem.* **1993**, *14*, 1429.

(13) Stechel, E. B., Ed. In *Domain Based Parallelism and Problem Decomposition in Computational Science and Engineering;* Keyes, D. R., Saad, Y., Truhlar, D. G., Eds.; SIAM: Philadelphia, PA, 1995.

(14) Baker, J.; Pulay, P. *J. Comput. Chem.* **2002**, *23*, 1150−1156.

(15) Nakao, Y.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 6375.

(16) Haettig, C.; Hellweg, A.; Koehn, A. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1159−1169.

(17) Ishimura, K.; Pulay, P.; Nagase, S. *J. Comput. Chem.* **2006**, *27*, 407−413.

(18) Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Theory Comput.* **2007**, *3*, 1312.

(19) Raghavachari, K.; Anderson, J. B. *J. Phys. Chem.* **1996**, *100*, 12960.

(20) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914.

(21) Galli, G.; Parrinello, M. *Phys. Rev. Lett.* **1992**, *69*, 3547.

(22) Murphy, R. B.; Beachy, M.; Ringnalda, M.; Friesner, R. *J. Chem. Phys.* **1995**, *103*, 1481.

(23) Nielsen, I. M. B.; Janssen, C. L. *J. Chem. Theory. Comput.* **2007**, *3*, 71.

(24) Lee, C.; Yang, W. *J. Chem. Phys.* **1992**, *96*, 2408.

(25) Baroni, S.; Giannozzi, P. *Europhys. Lett.* **1992**, *17*, 547.

(26) Théry, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. C. *J. Comput. Chem.* **1994**, *15*, 269.

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

(28) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(29) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2002**, *119*, 3599.

(30) Christie, R. A.; Jordan, K. D. *Struct. Bonding (Berlin)* **2005**, *116*, 27.

(31) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

(32) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.

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

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

(35) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2006**, *433*, 182.

(36) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**.

(37) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

(38) Fedorov, D. G.; Kitaura, K. *J. Comput. Chem.* **2007**, *28*, 222.

(39) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagese, S. *J. Comput. Chem.* **2007**, *28*, 1476.

(40) Fedorov, D. G.; Ishida, T.; Uebayasi, M.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 2722.

(41) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *312*, 319.

(42) Sugiki, S.; Kurita, N.; Sengoku, Y.; Sekino, H. *Chem. Phys. Lett.* **2003**, *382*, 611.

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

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

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

(46) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.

(47) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1995**, *96*, 6796.

(48) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(49) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.

(50) Hankins, D.; Moskowitz, J. W.; Stillinger, F. H. *J. Chem. Phys.* **1970**, *53*, 4544.

(51) Pedulla, J. M.; Jordan, K. D. *Chem. Phys. Lett.* **1998**, *291*, 78.

(52) Mhin, B. J.; Kim, H. S.; Kim, H. S.; Yoon, C. W.; Kim. K. S. *Chem. Phys. Lett.* **1991**, *176*, 41.

(53) Lee, C.; Chen, H.; Fitzgerald, G. *J. Chem. Phys.* **1994**, *101*, 4472.

(54) Estrin, D. A.; Paglieri, L.; Corongiu, G.; Clementi, E. *J. Phys. Chem.* **1996**, *100*, 8701.

(55) Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys.* **2000**, *112*, 2063.

(56) Lee, H. M.; Suh, S. B.; Lee, J. Y.; Tarakeshwar, P.; Kim, K. S. *J. Chem. Phys.* **2000**, *112*, 9759.

(57) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493.

(58) Losada, M.; Leutwyler, S. *J. Chem. Phys.* **2002**, *117*, 2003.

(59) Krishnan, R.; Pople, J. A. *Int. J. Quantum Chem.* **1978**, *14*, 91.

(60) Krishnan, R.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 4244.

(61) Frisch, M. J.; Krishnan, R.; Pople, J. A. *Chem. Phys. Lett.* **1980**, *75*, 66.

(62) Adams, G. F.; Bent, G. D.; Bartlett, R. J. In *Potential Energy Surfaces and Dynamics Calculations;* Truhlar, D. G., Ed.; Plenum: New York, 1981; p 133.

(63) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Robb, G. E. S. 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.; S. Clifford; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; P. Piskorz; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; M. A. Al-Laham; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; P. M. W. Gill; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03−version c01* eds.; Gaussian Inc.: Wallingford, CT, 2004.

(64) Halkier, A.; Koch, H.; Jorgensen, P.; Christiansen, O.; Nielsen, I. M. B.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 150.

(65) Nielsen, I. M. B.; Seidl, E. T.; Janssen, C. L. *J. Chem. Phys.* **1999**, *110*, 9435.

(66) Xantheas, S. S.; Aprà, E. *J. Chem. Phys.* **2004**, *120*, 823.

(67) Fanourgakis, G. S.; Aprà, E.; Xantheas, S. S. *J. Chem. Phys.* **2004**, *121*, 2655.

(68) Su, J. T.; Xu, X.; Goddard, W. A., III *J. Phys. Chem. A* **2004**, *108*, 10518.

(69) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493.

(70) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2002**, *2*, 1009.

CT700183Y

Journal of Chemical Theory and Computation

Revisiting the S_1/S_0 Degeneracy Space along the Exocyclic Methylene Twist Motion of Fulvene through a Two-Step Procedure

Masato Sumita and Kazuya Saito*

Department of Chemistry, Graduate School of Pure and Applied Sciences, University of Tsukuba, Tsukuba 305-8571, Japan

Received July 31, 2007

Abstract: We have characterized the degeneracy space (DS) between the ground (S_0) state and the first excited (S_1) state along the exocyclic methylene twist motion of fulvene, using our calculation strategy, i.e., a two-step procedure with CASSCF. The origin of the "cancellation error" on locating degeneracy points under geometrical constraints is analyzed, leading to a method to assess adequacy of the strategy. According to our estimation, these S_1/S_0 DPs are optimized for energy within 2.0×10^{-3} E_h Å^{-1} (the value of root-mean-square). From the obtained S_1/S_0 DS, we provide some information about the exocyclic methylene rotation by 180°.

1. Introduction

Recent theoretical calculations elucidated the importance of the conical intersections which are the real state crossing between the same spin multiplicity states.[1,2] A degeneracy point (DP), which is an apex of a conical intersection, is not an isolated point but consecutive space (see the next section about the details). The method to locate stationary DP (e.g., the lowest energy degeneracy point: LEDP) has been already established.[3,4] However, some theoretical calculations indicated the importance of exploring the degeneracy space (DS).[5−7] Hence, the method to explore the DS as a function of an arbitrary internal coordinate of the molecule is desired. Some methods characterizing the DS along an arbitrary internal coordinate of molecules have been reported. In the method based on Lagrange multipliers for optimization in the DS,[3] the determination of the section of the DS along a variable is possible.[8] In the projected gradient method,[4] when one uses the method with a geometric constrain beyond symmetry, the point at which energies are not degenerated is located. This undesirable result is called a "cancellation error'' which has been discussed, and some methods to circumvent the problem have been proposed.[7,9−12] According to these discussions, the origin of cancellation error is due to the loss of the orthogonality between a degeneracy lifting space and its complement space. We will however show that it is not the case.

We have circumvented the cancellation error by a two-step procedure.[9,10,13] We however did not assess how well energy was minimized with the two-step procedure. The goal of this paper is to clarify how well energy is minimized using the procedure and what condition is required for the procedure. To this end, we selected fulvene as a calculation target.

Fulvene is known as one of the isomers of benzene and a product of its photoisomerization.[14−17] The radiationless decay from the first excited (S_1) state in fulvene is observed.[18−21] Theoretically, this radiationless decay can be explained by the existence of some DPs.[12,22−24] These theoretical results suggested the possibility of the exocyclic methylene rotation by 180°. On the other hand, cis−trans photoisomerization is experimentally observed in the fulvene derivative. The photoisomerization of *E*−*Z*-2-*tert*-butyl-9-(2,2,2-triphenylthylidene)fluorene is recently observed experimentally.[25] This means that fulvene is useful as photoswitches if suitable substitutions are selected. To select the suitable substitutions, it is necessary to know the condition that makes it possible for the exocyclic methylene to rotate by 180°. Bearpark et al. suggested that the 0−0 excitation to S_1 is needed for the rotation.[22] In this paper, considering the existence of the S_1/S_0 DPs, we additionally discuss the condition of the exocyclic methylene rotation by 180°. Bearpark et al.[12] have already revealed that three S_1/S_0 DPs (DP$_{\text{planar}}$: C_{2v} planar structure, DP$_{63}$: exocyclic methylene is rotated by about 63° with C_2, and DP$_{\text{perp}}$: exocyclic

* Corresponding author e-mail: kazuya@chem.tsukuba.ac.jp.

10.1021/ct700190m CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/09/2007

methylene is perpendicular to a five-membered ring with C_{2v}) exist in the same S_1/S_0 DS which is predicted to be chemically relevant to cis−trans photoisomerization in contradiction to the suggestion by Deeb et al.[26] However, we have some questions about the previous mapping S_1/S_0 DS.[12] Here we will give more reliable results in the geometry with better energy degeneracy.

In section 2, we analyze the origin of the "cancellation error'' and suggest the method to assess the validity upon using our computational strategy.[9,10,13] In section 4, picking up the exocyclic methylene rotation of fulvene, we will show the valid condition in applying the two-step procedure based on section 2. Utilizing the procedure, the possibility of the methylene rotation in a fulvene molecule by 180° is discussed.

2. Theoretical Discussion

To describe the conical intersection, an apex of which is a DP, two coordinates are needed.[1,27] One is a gradient difference vector (GD)

$$\boldsymbol{g} = \nabla(E_1 - E_0) \tag{1}$$

and the other is a derivative coupling vector (DC)

$$\boldsymbol{h} = \langle \Psi_1 | \nabla \Psi_0 \rangle \tag{2}$$

In eqs 1 and 2, the gradient ∇ is a vector operator in nuclear coordinates. Ψ_1 and Ψ_0 are wave functions of the upper and lower states, respectively. Their energies are denoted as E_1 and E_0. The pair ($\boldsymbol{g}$, $\boldsymbol{h}$) is usually called a branching plane or a g-h plane.[1,27] In the complement orthogonal space to the branching plane, the degeneracy is preserved. In this paper, we refer to this complement space as a degeneracy space (DS) which is sometimes called a conical intersection hyperline or seam.[1,27] The DS is $(n - 2)$-dimensional space for two states, where n is the number of molecular internal degrees of freedom. We denote unit vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

$$\boldsymbol{x}_1 = \frac{\boldsymbol{g}}{|\boldsymbol{g}|}, \quad \boldsymbol{x}_2 = \frac{\boldsymbol{h}}{|\boldsymbol{h}|} \tag{3}$$

on the branching plane and $(n - 2)$-dimension internal coordinates orthogonal to the branching plane as $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, ..., $\boldsymbol{x}_n$. $\boldsymbol{x}_i$ ($i = 3, 4...n$) is referred to as intersection adapted coordinates.[27] Intersection adapted coordinates are different from nonredundant internal coordinates because each of the $\boldsymbol{x}_i$ ($i = 1, 2...n$) is represented as a linear combination of some variables like bond lengths, bond angles, and/or dihedral angles. To locate the lowest energy degeneracy point (LEDP) in DS, some optimization methods have been developed.[3,4,28] The projected gradient method[4] is extensively used. If this method is used together with a geometric constraint beyond molecular symmetry, however, a point at which the energy of two states are not degenerated is finally reached. We have pointed out that this error is due to constraining the variables that have components in the branching plane.[9] In the following discussion, we show that the error is due to constraining the variables that have components in both the branching plane and the intersection adapted coordinates.

In the projected gradient method, the following gradient is used

$$\boldsymbol{g}^{\mathrm{CIO}} = P\nabla E_1 + 2(E_1 - E_0)\nabla(E_1 - E_0) \tag{4}$$

where P is the projection operator onto the $(n - 2)$-dimensional intersection adapted coordinates. That is to say, P deducts DC and GD from ∇E_1. Here, we write the GD in a non-normalized form for simplicity although the GD in eq 4 is practically coded in a normalized form, $\boldsymbol{x}_1$.

Hereafter, we regard E_1 as a function of internal molecular coordinates, υ_i ($i = 1, 2, ..., n$), and define $\boldsymbol{e}_i$ as a unit vector in the direction of the displacement of υ_i. $\boldsymbol{e}_i$ must be orthogonal to each other. For instance, $\boldsymbol{e}_i$ can be obtained by orthogonalizing the unit vector of a physically significant set like bond lengths, bond angles, and dihedral angles.[29] Then, ∇E_1 can be represented by derivatives with respect to υ_i ($i = 1, 2, ..., n$).

$$\nabla E_1 = \frac{\partial E_1}{\partial \upsilon_1}\boldsymbol{e}_1 + \frac{\partial E_1}{\partial \upsilon_2}\boldsymbol{e}_2 + ... + \frac{\partial E_n}{\partial \upsilon_n}\boldsymbol{e}_n \tag{5}$$

For convenience, we classify the components of ∇E_1 into four groups.

$$\nabla E_1 = \frac{\partial E_1}{\partial \boldsymbol{\upsilon}_L}\boldsymbol{e}_L{}^{\mathrm{T}} + \frac{\partial E_1}{\partial \boldsymbol{\upsilon}_M}\boldsymbol{e}_M{}^{\mathrm{T}} + \frac{\partial E_1}{\partial \boldsymbol{\upsilon}_S}\boldsymbol{e}_S{}^{\mathrm{T}} + \frac{\partial E_1}{\partial \boldsymbol{\upsilon}_P}\boldsymbol{e}_P{}^{\mathrm{T}}$$

$$= \sum_h \frac{\partial E_1}{\partial \upsilon_{L,h}}\boldsymbol{e}_{L,h} + \sum_i \frac{\partial E_1}{\partial \upsilon_{M,i}}\boldsymbol{e}_{M,i} + \sum_j \frac{\partial E_1}{\partial \upsilon_{S,j}}\boldsymbol{e}_{S,j} + \sum_k \frac{\partial E_1}{\partial \upsilon_{P,k}}\boldsymbol{e}_{P,k} \tag{6}$$

where $\boldsymbol{\upsilon}_L$ is the group of the components which has no overlap with the branching plane. Both $\boldsymbol{\upsilon}_M$ and $\boldsymbol{\upsilon}_S$ are the groups of the components having overlap with the branching plane, but $\boldsymbol{\upsilon}_S$ is the variable that is constrained. On the other hand, $\boldsymbol{\upsilon}_P$ is the group of the components that lie within the branching plane. Corresponding unit vectors are denoted by $\boldsymbol{e}_L$, $\boldsymbol{e}_M$, $\boldsymbol{e}_S$, and $\boldsymbol{e}_P$ and distinguished by an additional subscript. After applying P, eq 6 becomes

$$P\nabla E_1 = \sum_h \frac{\partial E_1}{\partial \upsilon_{L,h}}\boldsymbol{e}_{L,h} + \sum_i c_{M,i}\frac{\partial E_1}{\partial \upsilon_{M,i}}\boldsymbol{e}_{M,i} + \sum_j c_{S,j}\frac{\partial E_1}{\partial \upsilon_{S,j}}\boldsymbol{e}_{S,j} \tag{7}$$

where coefficients, $c_{M,i}$ and $c_{S,j}$, satisfy

$$\begin{aligned} c_{M,i} &= 1 - \boldsymbol{x}_1 \cdot \boldsymbol{e}_{M,i} - 1 - \boldsymbol{x}_2 \cdot \boldsymbol{e}_{M,i} \\ &= 1 - c'_{M,i} \end{aligned} \tag{8a}$$

$$\begin{aligned} c_{S,j} &= 1 - \boldsymbol{x}_1 \cdot \boldsymbol{e}_{S,j} - 1 - \boldsymbol{x}_2 \cdot \boldsymbol{e}_{S,j} \\ &= 1 - c'_{S,j} \end{aligned} \tag{8b}$$

The branching plane component should be represented by the deducted component. Then we write the component of the second term in eq 4 as

$$\nabla(E_1 - E_0) = \sum_i c'_{M,i} \frac{\partial E_1}{\partial \upsilon_{M,i}} \boldsymbol{e}_{M,i} + \sum_j c'_{S,j} \frac{\partial E_1}{\partial \upsilon_{S,j}} \boldsymbol{e}_{S,j} + \sum_k \frac{\partial E_1}{\partial \upsilon_{P,k}} \boldsymbol{e}_{P,k} \quad (9)$$

Equation 4 then becomes

$$\boldsymbol{g}^{\mathrm{CIO}} = \sum_h \left(\frac{\partial E_1}{\partial \upsilon_{L,h}}\right) \boldsymbol{e}_{L,k} + \sum_i \left(c_{M,i}\left(\frac{\partial E_1}{\partial \upsilon_{M,i}}\right) + 2(E_1 - E_0)c'_{M,i}\left(\frac{\partial E_1}{\partial \upsilon_{M,i}}\right)\right)\boldsymbol{e}_{M,i} + \sum_j \left(c_{S,j}\left(\frac{\partial E_1}{\partial \upsilon_{S,j}}\right) + 2(E_1 - E_0)c'_{S,j}\left(\frac{\partial E_1}{\partial \upsilon_{S,j}}\right)\right)\boldsymbol{e}_{S,j} + \sum_k 2(E_1 - E_0)\left(\frac{\partial E_1}{\partial \upsilon_{P,k}}\right)\boldsymbol{e}_{P,k} \quad (10)$$

Although the third summation term is eliminated for geometric constraint, we keep this term for clear discussion. The following condition is also implicitly imposed because of the orthogonality between the intersection adapted coordinates and branching plane:

$$P\nabla E_1 \cdot \nabla(E_1 - E_0) = \sum_i c_{M,i}c'_{M,i}\left(\frac{\partial E_1}{\partial \upsilon_{M,i}}\right)^2 + \sum_j c_{S,j}c'_{S,j}\left(\frac{\partial E_1}{\partial \upsilon_{S,j}}\right)^2 = 0 \quad (11)$$

According to eq 10, the convergence condition then reads

$$\frac{\partial E_1}{\partial \upsilon_{L,h}} = 0 \quad (12a)$$

$$c_{M,i}\frac{\partial E_1}{\partial \upsilon_{M,i}} + 2(E_1 - E_0)c'_{M,i}\frac{\partial E_1}{\partial \upsilon_{M,i}} = 0 \quad (12b)$$

$$c_{S,j}\frac{\partial E_1}{\partial \upsilon_{S,j}} + 2(E_1 - E_0)c'_{S,j}\frac{\partial E_1}{\partial \upsilon_{S,j}} = C_{S,j} \quad (12c)$$

$$2(E_1 - E_0)\frac{\partial E_1}{\partial \upsilon_{P,k}} = 0 \quad (12d)$$

Here, $C_{S,j}$ is finite. Equation 12(a) shows that optimization will be successful if the variables that have no overlap with the branching plane are employed. As for eq 12(d), two situations are possible. One is $E_1 - E_0 = 0$ and the other is $\partial E_1/\partial \upsilon_{P,k} = 0$. The former condition is, however, ruled out by the following reason: Multiplying eq 12(b) by $c_{M,i}(\partial E_1/\partial \upsilon_{M,i})$ and using eq 11

$$\sum_j c_{S,j}c'_{S,j}\left(\frac{\partial E_1}{\partial \upsilon_{S,j}}\right)^2 = 2(E_1 - E_0)\sum_i {c'_{M,i}}^2\left(\frac{\partial E_1}{\partial \upsilon_{M,i}}\right)^2 \quad (13)$$

is obtained. Apart from special cases (e.g., the value of $\upsilon_{S,j}$ corresponds to that of a LEDP), $c_{S,j}c'_{S,j}(\partial E_1/\partial \upsilon_{S,j})^2$ is not zero from eq 12c. The right-hand side of eq 13 is not zero, accordingly. Namely, the optimization converges to the point where two states do not degenerate (i.e., $E_1 \neq E_0$). This is really a "cancellation error". If either $c_{S,j}$ or $c'_{S,j}$ is zero, then the cancellation error does not occur because there are no dependences between $\partial E_1/\partial \upsilon_{M,i}$ and $\partial E_1/\partial \upsilon_{S,j}$ by eq 11. If both $c_{S,j}$ and $c'_{S,j}$ are not zero, then the cancellation error occurs. Therefore, in contradiction to the previous suggestion (the orthogonality between the first and second term in eq 4 is lost due to the constraint), to keep the orthogonal condition [eq 11], the first term offsets the second term in eq 4.

Table 1. Values of the Difference ($E_1 - E_0$) (in E_h) and the Gradient, Eq 4 to θ, and $0.5\sqrt{2(E_1 - E_0)}|C_{S,\theta}|$ (in E_h Å^{-1}) along θ in the First Step[a]

θ (deg)	($E_1 - E_0$)	$C_{S,\theta}$	$0.5\sqrt{2(E_1 - E_0)}\lvert C_{S,\theta}\rvert$	RMS
0	0.00000	0.00000	0.00000	0.00000
5	0.00002	−0.01530	0.00005	0.00030
10	0.00008	−0.00890	0.00005	0.00058
15	0.00016	−0.01294	0.00012	0.00085
20	0.00028	−0.01647	0.00019	0.00108
25	0.00041	−0.01928	0.00028	0.00127
30	0.00053	−0.02115	0.00035	0.00140
35	0.00064	−0.02185	0.00039	0.00145
40	0.00070	−0.02119	0.00040	0.00141
45	0.00070	−0.01902	0.00036	0.00127
50	0.00061	−0.01530	0.00027	0.00102
55	0.00042	−0.01018	0.00015	0.00068
60	0.00017	−0.00405	0.00004	0.00027
63.1	0.00000	0.00000	0.00000	0.00000
65	0.00010	0.00235	0.00002	0.00016
70	0.00031	0.00791	0.00010	0.00053
75	0.00038	0.01125	0.00016	0.00074
80	0.00028	0.01103	0.00013	0.00072
85	0.00009	0.00686	0.00005	0.00044
90	0.00000	0.00000	0.00000	0.00000

[a] The RMS values of the projected gradient which is obtained after the second step is also listed.

Recently, this cancellation error has been circumvented by several methods.[7,9−12] Migani et al.[7] circumvented it by scaling the second term in eq 4 with a factor of 100. Yamazaki et al.[11] circumvented it by orthogonalizing the internal coordinates of molecules. With the gradient of which the constraint is applied before the projection of ∇E_1 onto the intersection adapted coordinates, Bearpark et al.[12] have succeeded to map the S_1/S_0 DS along the exocyclic methylene rotation of fulvene with a maximum energy gap of 0.4 kcal mol^{-1}. It is, however, noteworthy that the points at which a maximum energy gap is approximately 0.4 kcal mol^{-1} (see Table 1) can be located by using the default gradient (our first step). That is, there is no difference in effect between the default gradient[4] and the modified gradient.[12]

On the other hand, in our easy computational strategy, after optimization using eq 4 (i.e., converging to the geometry satisfying eq 12), we carried out the geometry optimization using only the second term in eq 4.[9,10,13] We have used this computational strategy without estimating how well energy is minimized within the intersection adapted coordinates. Here we try to assess the validity of the strategy. Multiplying eq 12b by $c_{M,j}(\partial E_1/\partial \upsilon_{M,i})$ and using eq 11, we obtain

$$\sum_i {c_{M,i}}^2\left(\frac{\partial E_1}{\partial \upsilon_{M,i}}\right)^2 = 2(E_1 - E_0)\sum_j c_{S,j}c'_{S,j}\left(\frac{\partial E_1}{\partial \upsilon_{S,j}}\right)^2 \quad (14)$$

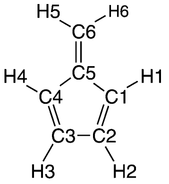


Figure 1. Atom numbering using throughout this paper.

This equation indicates the variables which have overlap with both the branching plane and the intersection adapted coordinates cannot work as an independent variable for optimization due to geometric constraint. In turn, the second step optimization is successfully limited in effect within the intersection adapted coordinates if eq 14 is small enough. The two-step procedure is, in this sense, not for improving the energy degeneracy but for the better geometry in the DS. Validity (or limitation) of the strategy can be assessed by using eq 14 as will be shown later.

3. Computational Details

All calculations in this paper were carried out using the CASSCF method implemented in GAUSSIAN 98[30] with the correlation-consistent polarized valence double-zeta (cc-pVDZ) basis set. An active space of six electrons in six orbitals was used, corresponding to π orbitals. CASSCF were carried out using the S_1/S_0 state-averaged orbital, with the two states weighted equally.

To characterize S_1/S_0 DS, we carried out two-step optimizations described in the previous section. In the first step, we used eq 4 as gradient until the square root of eq 14 becomes sufficiently small as will be shown in the next section. In the second step, we used only the second term in eq 4.

Starting from C_{2v} planar structures, the S_1/S_0 DS was scanned in C_2 symmetry along the exocyclic methylene twist motion with a step size of 5° up to C_{2v} twisted structures.

Our calculation is not definitive because the CASSCF does not take into account effects of dynamical electronic correlation. However, the behavior we have predicted in this paper would not be affected qualitatively by it.

4. Results and Discussion

The atomic numbering is shown in Figure 1. Hereafter, θ denotes the twist angle of the exocyclic methylene. In Figure 2, we show the example of the two-step procedure locating DP at $\theta = 45°$.

In eq 14, we have shown that the variables, $\boldsymbol{v}_M$, that have overlap with both the intersection adapted coordinates and the branching plane are dependent on the constrained variables that also have overlap with these two spaces. From eq 14, the square root of the gradient for $\boldsymbol{v}_M$ in the intersection adapted coordinates is given by

$$\sqrt{\sum_i c_{M,i}{}^2\left(\frac{\partial E_1}{\partial v_{M,i}}\right)^2} = \sqrt{2(E_1 - E_0)\sum_j c_{S,j}c'_{S,j}\left(\frac{\partial E_1}{\partial v_{S,j}}\right)^2} \quad (15)$$

According to eq 15, how geometries are well optimized in intersection adopted coordinates depends on the magnitude of $E_1 - E_0$ and the gradient with respect to constrained variables ($\partial E_1/\partial \boldsymbol{v}_S$) in the first step [using the default gradient, eq 4]. The value of ($\partial E_1/\partial v_{S,j}$) can roughly be estimated by $C_{S,j}$ in eq 12c. As $C_{S,j}$ includes the normalization factor of the GD, the values of $C_{S,j}$ are larger than ($\partial E_1/\partial v_{S,j}$). In the system we targeted, only one variable, θ, is constrained. It is known that mutual transformation between forces represented by Cartesian coordinates and by nonredundant internal coordinates is possible.[31] Furthermore, the physically significant set can be written by the linear combination of $\boldsymbol{e}_i$. Therefore, the right-hand side of eq 15 can be written by using θ.

$$\sqrt{\sum_i c_{M,i}{}^2\left(\frac{\partial E_1}{\partial v_{M,i}}\right)^2} = \sqrt{2(E_1 - E_0)c_{S,\theta}c'_{S,\theta}}\left|\frac{\partial E_1}{\partial v_{S,\theta}}\right| \le \sqrt{2(E_1 - E_0)c_{S,\theta}c'_{S,\theta}}C_{S,\theta} \quad (16)$$

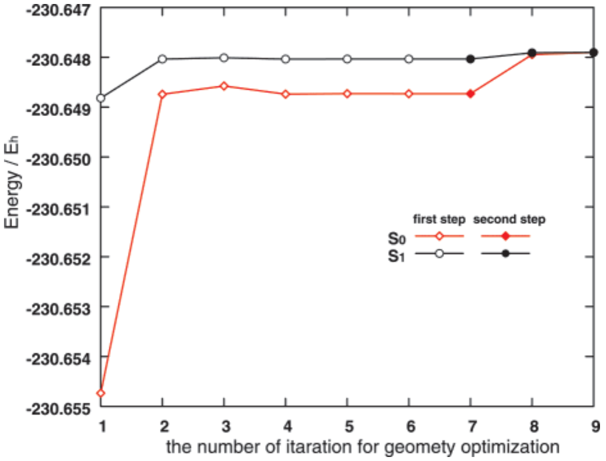


Figure 2. The example of the two-step procedure in locating DP at $\theta = 45°$. The starting structure was produced by replacing the value of θ of the DP at $\theta = 40°$ by 45°. Open symbols (diamond and circle) indicate the first step iteration. Filled symbols (diamond and circle) indicate the second step iteration. At iteration number 6, the first step (using the default gradient $\boldsymbol{g}^{\mathrm{CIO}}$ (eq 4)) was completed, whereas the second step (using the second term of $\boldsymbol{g}^{\mathrm{CIO}}$ (eq 4)) started from iteration number 7.

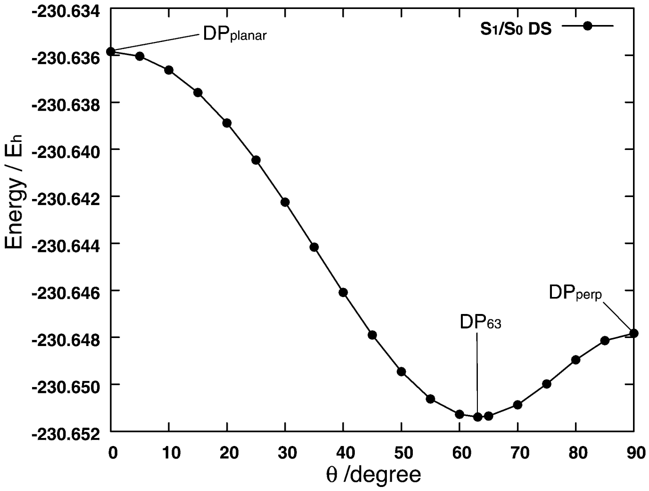


Figure 3. The result of the S_1/S_0 DS along θ.

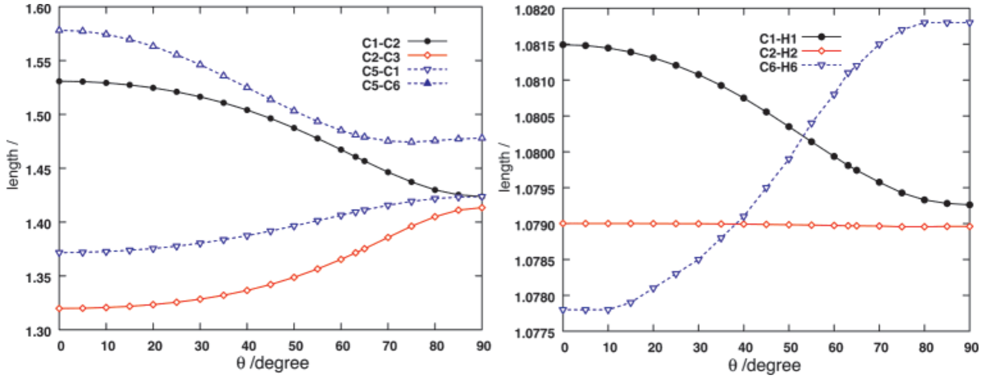


Figure 4. Geometric change along S_1/S_0 DS. (a) carbon−carbon bond lengths and (b) carbon−hydrogen bond lengths.

The values of the $\sqrt{(E_1-E_0)}$, which is the square root of the difference between energies of the S_1 and S_0 and $C_{S,\theta}$ that is the value of eq 4 as the gradient along θ, are given in Table 1. Furthermore, the upper bound of $\sqrt{c_{S,\theta}c'_{S,\theta}}$ can also be estimated by the inequality between arithmetic and geometric means, i.e., $\sqrt{c_{S,\theta}c'_{S,\theta}} \leq 0.5$. The degree of optimization in the variables which overlap with both intersection adapted coordinates and the branching plane can therefore be estimated as $0.5\sqrt{2(E_1-E_0)}C_{S,\theta}$ approximately. From Table 1, the gradient of E_1 with respect to the variables which correspond to v_M is approximately 0.0004 (in E_h Å^{-1}). From our experience, this magnitude is small enough to reoptimize from each point obtained in the first step to locate the S_1/S_0 DP using the second term of eq 4. We show the root-mean-square (RMS in Cartesian coordinate) of projected gradient on E_1 [the first term in eq 4] whose component of exocyclic methylene rotation is given in Table 1. According to the values of RMS of Table 1, the geometry of the finally obtained DP is optimized within 1.5×10^{-3} E_h Å^{-1} at worst. These RMS values are comparable to those of the residual gradient optimized “loosely” by GAUSSIAN 98. There is one more important condition for validity of the two-step procedure. The tendency of change of the value of the RMS indicates a similar change of the value of (16) along θ. This means the final geometry may be in the same intersection adapted coordinates of the geometry which is obtained in the first step. From our experiences, if the tendency of the change of the final RMS is different from that of eq 16, resultant DS would not be meaningful. In Figure 2, we show the example of the two-step procedure locating DP at $\theta = 45°$.

Now, it is in order to see some details of the characterized DS. The S_1/S_0 DS characterized S_1/S_0 using the above strategy is shown in Figure 3. A recent second-order derivative calulation in the S_1/S_0 DS[24] has revealed that DP_{planar} and DP_{perp} are second- and first-order saddle points, and DP_{63} is almost the global minium on the S_1 excited state and S_1/S_0 DS though its energy is slightly lowerd by pyramidalization.[32] Our result is favorably compared with the second derivative calculation. The energies of the two states agreed within 10^{-5} E_h for all the DPs located. Starting from DP_{planar}, we have characterized the S_1/S_0 DS along θ up to DP_{perp}. This result also tells us that θ is the variable which has overlap with both the intersection adapted coordinates and the branching plane. Unless so, the first step optimization should converge to DP.

The origin of degeneracy of DP_{planar} and DP_{perp} is different. In DP_{planar}, the degeneracy occurs by elongating the exocyclic double bond and enhanced allylic character. On the other hand, in DP_{perp},[22] the degeneracy stems from the D_1/D_0 symmetry required conical intersection of the cyclopentadienyl radical:[33,34] The bonds that compose the five-membered ring become more similar to each other. Indeed, the bond lengths of C1−C2, C5−C1, and C2−C3 become about 1.4 Å equally. In spite of the different origin of the DP, Figure 4 shows that the electronic structure is continuously changed from DP_{planar} to DP_{perp}. This demonstrates that DP_{planar} and DP_{perp} are in the same DS.

The behavior of the exocyclic double bond C5−C6 is very interesting. We expected that the tendency of the geometric change of C5−C6 is changed in the vicinity of DP_{63} corresponding to the global minimum on the S_1 state. However, around DP_{63} (i.e., around $\theta = 60°$), there are no particular changes. This implies that the electronic structure is not changed around DP_{63}. Instead, the tendency of the geometric change of C5−C6 is changed around DP_{75}. Hence, we can imagine that the DPs between $\theta = 0°$ and $\theta = 75°$ will be photochemically discriminated from the DPs between $\theta = 80°$ and $\theta = 90°$. To clarify the final product via the S_1 state, we have performed S_0 geometry optimizations using a state-averaged orbital from structures near DP_{63}, DP_{75}, and DP_{80}. Starting structures were generated by distorting the DP geometries in the direction of GD. These results indicate that the product whose exocyclic methylene is rotated by 180° is available from DP_{80} but not from DP_{63} and DP_{75}. Therefore, if the S_1 excited fulvene can reach the DPs between $\theta = 80°$ and $\theta = 90°$, then the exocyclic methylene rotation by 180° is possible. If DPs in this area are stabilized by the proper substitution so that the S_1 excited fulvene can reach this area, then cis−trans photoisomerization will become possible. In the dibenzofulvene system whose $E-Z$ photoisomerization is observed recently,[25] adding the benzene to fulvene may give rise to the stabilization of the DPs between $\theta = 80°$ and $\theta = 90°$.

5. Conclusion

We have shown that the cancellation error is due to the constraint of the variables that have components in both the branching plane and the intersection adapted coordinates. Accordingly, the valid condition for the two-step procedure is limited. Taking into account the limitation, we have characterized the S_1/S_0 DS along the exocyclic methylene rotation coordinate of fulvene within 1.5×10^{-3} E_h Å^{-1} in energy at worst.

Our calculation, which we have shown in this paper, is limited to C_2 symmetry. Though systems which have no symmetry like ref 25 should be explored, the following conclusion would be worthy to be noted. The photophysical/photochemical behavior changes in the continuous DS. The DPs where the photochemical property changes are not the saddle point on the S_1/S_0 DS within C_2 symmetry. That is, the product obtained via S_1/S_0 DPs in the vicinity of DP_{63} does not change. It is difficult for the exocyclic methylene to rotate by 180°, when the S_1 excited fulvene transits to S_0 via DPs between DP_{planar} and DP_{75}. However, in DPs between DP_{80} and DP_{perp}, the exocyclic methylene rotation is expected. Therefore, photochemically, $DP_{80-perp}$ may be discriminated from $DP_{planar-75}$.

Acknowledgment. We are grateful to Dr. Ohgi Takahashi for a stimulating discussion. We sincerely acknowledge the reviewers of this manuscript for valuable suggestions and comments.

Supporting Information Available: Cartesian coordinates of DPs geometries discussed in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bernardi, F.; Olvucci, M.; Robb, M. A. *Chem Soc. Rev.* **1996**, *25*, 321−328.

(2) Migani, A.; Olivucci, M. Conical Intersection and Organic Reaction mechanisms. In *Conical Intersections: Electronic Structure, Dynamics and Spectroscopy, Advance Series in Physical Chemistry*; Comcke, W., Yarkony, D. R., Köppel, H., Eds.; World Scientific: Singapore, 2004; Vol. 15, pp 271−320.

(3) Manaa, M. R.; Yarkony, D. R. *J. Am. Chem. Soc.* **1994**, *116*, 11444−11448.

(4) Bearpark, M. J.; Robb, M. A.; Schlegel, H. B. *Chem. Phys. Lett.* **1994**, *223*, 269−274.

(5) Palmer, I. J.; Ragazos, I. N.; Bernardi, F.; Olivucci, M.; Robb, M. A. *J. Am. Chem. Soc.* **1993**, *115*, 673−682.

(6) Venturini, A.; Vreven, T.; Bernardi, F.; Olivucci, M.; Robb, M. A. *Organometallics* **1995**, *14*, 4953−4956.

(7) Migani, A.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **2003**, *125*, 2804−2808.

(8) Yarkony, D. R. *J. Phys. Chem. A* **2004**, *108*, 3200−3205.

(9) Takahashi, O.; Sumita, M. *J. Chem. Phys.* **2004**, *121*, 7030−7031.

(10) Takahashi, O.; Sumita, M. *J. Mol. Struct. THEOCHEM* **2005**, *731*, 173−175.

(11) Yamazaki, S.; Kato, S. *J. Chem. Phys.* **2005**, *123*, 114510−13.

(12) Bearpark, M. J.; Blancafort, L.; Paterson, M. J. *Mol. Phys.* **2006**, *104*, 1033−1038.

(13) Sumita, M.; Saito, K. *Chem. Phys. Lett.* **2006**, *424*, 374−378.

(14) Shindo, K.; Lipsky, S. *J. Chem. Phys.* **1996**, *45*, 2292−2297.

(15) Foote, J. K.; Mallon, M. H.; Pitts, J. N., Jr. *J. Am. Chem. Soc.* **1966**, *88*, 3698−3702.

(16) Wilzbach, K. E.; Harkness, A. L.; Kaplan, L. *J. Am. Chem. Soc.* **1968**, *90*, 1116−1118.

(17) Kaplan, L.; Wilzbach, K. E. *J. Am. Chem. Soc.* **1968**, *90*, 3291−3292.

(18) Kent, J. E.; Harman, P. J.; O'Dwyer, M. F. *J. Phys. Chem.* **1981**, *85*, 2726−2730.

(19) Harman, P. J.; Kent, J. E.; O'Dwyer, M. F.; Smith, M. H. *Aust. J. Chem.* **1979**, *32*, 2579−2587.

(20) Domaille, P. J.; Kent, J. E.; O'Dwyer, M. F. *Chem. Phys.* **1974**, *6*, 66−75.

(21) Brown, R. D.; Domaille, P. J.; Kent, J. E. *Aust. J. Chem.* **1970**, *23*, 1707−1720.

(22) Bearpark, M. J.; Bernardi, F.; Olivucci, M.; Robb, M. A.; Smith, B. R. *J. Am. Chem. Soc.* **1996**, *118*, 5254−5260.

(23) Dreyer, J.; Klessinger, M. *J. Chem. Phys.* **1994**, *101*, 10655−10665.

(24) Paterson, M. J.; Bearpark, M. J.; Robb, M. A.; Blancafort, L. *J. Chem. Phys.* **2004**, *121*, 11562−11571.

(25) Barr, J. W.; Bell, T. W.; Catalano, V. J.; Cline, J. I.; Phillips, D. J.; Procupez, R. *J. Phys. Chem. A* **2005**, *109*, 11650−11654.

(26) Deeb, O.; Cogan, S.; Zilberg, S. *Chem. Phys.* **2006**, *325*, 251−256.

(27) Yarkony, D. R. Conical Intersection and Organic Reaction mechanisms. In *Conical Intersections: Electronic Structure, Dynamics and Spectroscopy, Advance Series in Physical Chemistry*; Comcke, W., Yarkony, D. R., Köppel, H., Eds.; World Scientific: Singapore, 2004; Vol. 15, pp 41−127.

(28) Dallos, M.; Lischka, H.; Shepard, R.; Yarkony, D. R.; Szalay, P. G. *J. Chem. Phys.* **2004**, *120*, 7330−7339.

(29) Lu, D.-H.; Zhao, M.; Truhlar, D. G. *J. Comput. Chem.* **1991**, *12*, 376−384.

(30) 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.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.;

Wong, M. W.; Andres, J. L. M.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98 (Revision A.11.3)*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(31) Schlegel, H. B. *Theor. Chim. Acta* **1984**, *66*, 333−340.

(32) Sicilia, F.; Bearpark,M. J.; Blancafort, L.; Robb, M. A. *Theor. Chem. Acc.* **2007**, *118*, 241−251.

(33) Borden, W. T.; Davidson, E. R. *J. Am. Chem. Soc.* **1979**, *101*, 3771−3775.

(34) Yu, L.; Cullin, D. W.; Williamson, J. M.; Miller, T. A. *J. Chem. Phys.* **1993**, *98*, 2682−2682.

CT700190M

JCTC Journal of Chemical Theory and Computation

Inductive and External Electric Field Effects in Pentacoordinated Phosphorus Compounds

Enrique Marcos, Ramon Crehuet,* and Josep M. Anglada*

Grup de Química Teòrica i Computacional, Departament de Química Orgànica Biològica, Institut d'Investigacions Químiques i Ambientals de Barcelona, IIQAB − CSIC, c/ Jordi Girona 18, E-08034 Barcelona, Spain

Received August 28, 2007

Abstract: Pentacoordination at phosphorus is associated with a nucleophilic displacement reaction at tetracoordinated phosphorus compounds and shows a great variability in what respects their geometrical and energetic features. By means of a systematic theoretical study on a series of elementary model compounds, we have analyzed the bonding features. The pentacoordinated phosphorus compounds are held together by dative bonds, and the geometry and stability depends on the inductive effects originated by different substitutes at phosphorus. We show also that an external electric field can modify the geometrical features and the reactivity of the nucleophilic substitution reactions. This issue may have great interest in biological reactions involving pentacoordinated phosphorus where the electric field originated by the folded protein could influence the catalytic process. We report also additional calculations on the geometry and NMR spectra on three triphenyl phosphonium ylide derivatives, and our results compare well with the experimental data.

Introduction

A detailed knowledge on the electronic nature of pentacoordination at phosphorus is of great interest in chemistry and biochemistry.[1−12] Pentacoordination at phosphorus is mainly associated with a nucleophilic displacement reaction at tetracoordinated phosphorus compounds, which is associated with cell signaling and energetics and many aspects of biosynthesis. These nucleophilic reactions occur in the so-called associative processes, which can follow a concerted pathway, with a trigonal bipyramid transition state, or an addition−elimination pathway, involving a pentacoordinated phosphorane intermediate.[7−9,11,13,14] These processes are important in chiral reactions. Those following a concerted pathway take place with inversion of configuration, but in pathways involving pentacoordinate intermediates, a Berry pseudorotation may occur, which could involve retention of configuration.[5,11] Pentacoordinated phosphorus intermediates are found, for instance, in the Wittig reaction,[15] in human α-thrombin inhibitors,[1] and as intermediates in the hydrolysis of phospholipids catalyzed by phospholipase D.[6] It may exist in phosphoryl transfer in GTP hydrolysis by RAS proteins[10] and as an intermediate in the phosphoryl transfer reaction catalyzed by a β-phosphoglucomutase,[2,16] although some controversy exists in the literature regarding the true nature of this intermediate.[3,4]

Pentacoordination at phosphorus occurs mainly in trigonal bipyramid structures, and it has been observed that the apical bond lengths show a great variability, which depends on several factors as the nature of the substitutes at P, the influence of hydrogen bonding or the charge around phosphorus.[1,10,11,17,18] It appears therefore that such variability would affect not only the stability of these compounds but also the transition states involving pentacoordination at phosphorus and consequently the reactivity. The factors affecting this variability are crucial for a complete understanding of nucleophilic displacement at phosphorus. They are still not well rationalized and are the main goal of this study.

Extensive theoretical studies have also been reported in the literature, which have provided valuable information regarding different aspects of the reaction mechanisms of

* Corresponding author e-mail: anglada@iiqab.csic.es (J.M.A.), rcsqtc@iiqab.csic.es (R.C.).

10.1021/ct700220z CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/05/2007

phosphate reactions, the importance and possible existence of pentacoordinated intermediates depending on the reaction conditions, and the effects of the solvent in the reactivity.[19−47] In this study we have focused our attention on the inductive effects affecting pentacoordination at phosphorus and its bonding features. To this end, we have considered, in the first stage, a series of model systems for which we have investigated the effect of different substitutes at phosphorus as well as the effect of polarization and the effect of an external electric field. In the second stage, we have also investigated a series of triphenylphosphonium ylide derivatives for which experimental data exist in the literature.

Computational Details

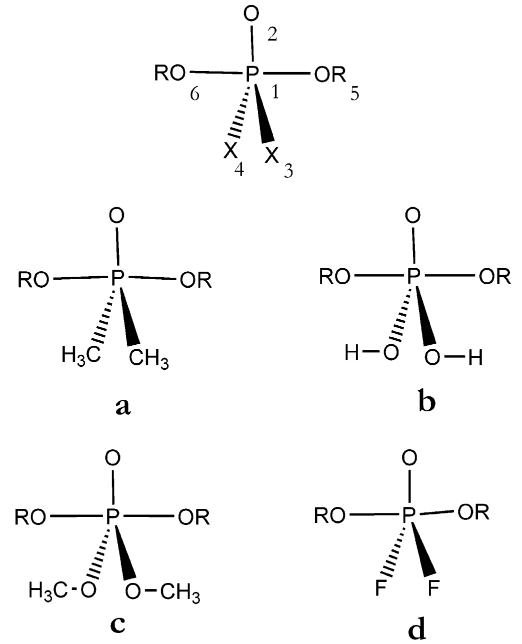
All geometry optimizations carried out in this work have been performed with the density functional *m*PW1PW91[48] employing the 6-31+G(d) basis set.[49] At this level of theory we have also calculated the harmonic vibrational frequencies to verify the nature of the corresponding stationary point (minima or transition state) and to provide the zero point vibrational energy (ZPE). The *m*PW1PW91 functional has been found to be adequate to describe systems with long-range interactions, especially with dative bonds.[50] Moreover, the reliability of this functional, with respect to the geometrical parameters, has also been checked by performing, for some test models, comprehensive test calculations employing the MP2[51−53] ab initio approach and using the 6-31+G(d), 6-311+G(d), and 6-311+G(3df, 3pd) basis sets. The results obtained compare quite well and are collected in the Supporting Information. Moreover, for the cases where reactivity has been considered, we have performed, for each transition state, intrinsic reaction coordinate calculations (IRC)[54−56] in order to ensure that the transition states connect the desired reactants and products. In the second step, the relative energies of the stationary points were corrected by performing single point energy calculations using the *m*PW1PW91 functional with the 6-311+G(3df,2p) basis set.[57] In addition, we have also checked the reliability of the activation and reaction energies by performing, for all stationary points of a given reaction (reaction **1b**, see below), additional single point energy calculations at the higher level of theory CCSD(T)/IB.[58−62] The results obtained at the *m*PW1PW91 and CCSD(T) level compare very well and are contained in the Supporting Information.

For the three triphenylphosphonium ylides considered, we have also computed the NMR spectra by performing B3LYP single-point calculations[63] at the optimized geometries, using the GIO method[64,65] and employing the 6-311+G(2d,p) basis set.[57]

The quantum chemical calculations carried out in this work were performed by using the Gaussian[66] program package, and the Molden program[67] was employed to visualize the geometric and electronic features.

The bonding features of the different systems considered were analyzed by employing the natural bond orbital (NBO) partition scheme by Weinhold and co-workers[68] and the atoms in molecules (AIM) theory by Bader.[69] The topological properties of wave functions were computed using the AIMPAC program package.[70]

Scheme 1 [a]



[a] RO = HO (**1**); CH_3O (**2**); HCOO (**3**); CF_3O (4).

Results and Discussion

The Model Systems $POX_2(RO)_2$ Pentacoordinated Compounds. One important point regarding the chemistry of pentacoordinated phosphorus compounds refers to the variability of the apical bond distances.[5,11] In order to analyze and rationalize this issue we have carried out a series of calculations on the $POX_2(RO)_2$ model systems. These pentacoordinated model systems have been depicted in Scheme 1 and possess a trigonal bipyramid structure. Here X are equatorial substitutes (X = CH_3 (**a**); HO (**b**); CH_3O (**c**); and F (**d**)) and RO are apical substitutes (RO = HO (**1**); CH_3O (**2**); HCOO (**3**); and CF_3O (**4**)). Along this work, the different models are labeled by a number as a prefix, according to the apical substitutes, followed by a letter as a suffix according to the equatorial substitutes. Thus, compound **1a** corresponds to $PO(CH_3)_2(OH)_2$, whereas compound **3c** corresponds to $PO(CH_3O)_2(HCOO)_2$ (see Scheme 1).

Please note also that in these model systems the charge of the system is −1, the two apical substitutes are identical, and that in **b** and **c** the two equatorial substitutes are oppositely oriented. Moreover, it is also worth reminding the reader that the donor character of the apical substitutes is HO > CH_3O > HCOO > CF_3O and the donor character of the equatorial substitutes is CH_3 > HO > CH_3O > F so that these series of model systems allow us to analyze combinations of electron withdrawing groups and electron donor groups on the phosphorus coordination. The most significant geometrical parameters of the optimized structures are displayed in Table 1, which also includes the tetracoordinated phosphoric acid H_3PO_4 for comparison. Figure 1 shows the dependence of the apical bond lengths with respect to the apical and equatorial substitutes at P. The Cartesian coordinates of each pentacoordinated model system are reported in the Supporting Information.

For H_3PO_4, Table 1 shows that our calculations predict the P···O bond length to be 1.476 Å and the three P···OH

Table 1. Optimized Bond Lengths (in Å) to Phosphorus in H_3PO_4 and the Pentacoordinated **1a**−**4d** Model Compounds

	substitutes				
compd	apical	equatorial	r(P−OR$_{apical}$)	r(P−X$_{equatorial}$)	r(P−O)
H_3PO_4				1.602	1.476
1a	OH	CH_3	1.814	1.849	1.545
1b	OH	OH	1.768	1.660	1.527
1c	OH	OCH_3	1.742	1.673	1.539
2a	OCH_3	CH_3	1.841	1.841	1.521
2b	OCH_3	OH	1.778	1.660	1.510
2c	OCH_3	CH_3	1.749	1.677	1.514
3a	OC(H)O	CH_3	1.908	1.832	1.507
3b	OC(H)O	OH	1.835	1.635	1.495
3c	OC(H)O	OCH_3	1.818	1.639	1.499
3d	OC(H)O	F	1.780	1.609	1.496
4a	OCF_3	CH_3	1.980	1.826	1.488
4b	OCF_3	OH	1.878	1.625	1.480
4c	OCF_3	OCH_3	1.882	1.616	1.484
4d	OCF_3	F	1.808	1.591	1.476

Table 2. Natural Occupation at Phosphorus, Stabilization Energies ($\Delta E(2)$ in kcal·mol^{-1}) Associated with the Most Important Donor−Acceptor Interaction Involving the Apical Bonds and the Natural Charges at Phosphorus (Q in e)[d]

	P natural occupation			stabilization energies			
compd	s	p	d	$\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ [a]	$\sigma_{P1X} \rightarrow \sigma^*_{P1O6}$ [b]	$\sigma_{P1O2} \rightarrow \sigma^*_{P1O6}$ [c]	Q_{NBO}(P)
1a	0.91	1.85	0.09	33.51	30.67	28.88	2.12
1b	0.77	1.59	0.11	29.80	22.66	21.59	2.51
1c	0.77	1.56	0.11	28.86	20.02	23.54	2.54
2a	0.91	1.81	0.08	31.86	39.14	27.90	2.17
2b	0.77	1.55	0.10	25.50	30.71	20.28	2.56
2c	0.76	1.51	0.10	24.27	24.41	16.98	2.59
3a	0.94	1.82	0.08	35.52	35.23	31.58	2.13
3b	0.77	1.57	0.10	29.04	24.04	23.34	2.53
3c	0.77	1.53	0.10	28.80	24.08	22.12	2.58
3d	0.76	1.49	0.11	24.88	20.59	24.90	2.62
4a	0.95	1.82	0.07	35.94	38.97	34.06	2.12
4b	0.77	1.57	0.10	29.43	24.43	26.59	2.53
4c	0.77	1.53	0.10	32.39	25.32	27.89	2.58
4d	0.75	1.50	0.11	28.48	22.90	28.43	2.62

[a] $\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ has the same value as $\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$. [b] The same interaction occurs from each σ_{PX} equatorial to each σ^*_{PO} apical bond. [c] $\sigma_{P1O2} \rightarrow \sigma^*_{P1O6}$ has the same value as $\sigma_{P1O2} \rightarrow \sigma^*_{P1O5}$. [d] Bond numbering is according Scheme 1.

bond lengths to be 1.602 Å. The nucleophilic addition of the HO anion to H_3PO_4 leads to the pentacoordinated compound **1b** and produces a lengthening of 0.051 Å in the P···O bond and of 0.058 Å in the equatorial P···OH bonds, compared with the P···O and the P···OH bond lengths in phosphoric acid, but the two apical P···OH bond distances (1.768 Å) are predicted to be much longer (see Table 1).

Regarding the remaining pentacoordinated model systems, the results of Table 1 and Figure 1 show a great variability of the P···OR$_{apical}$ bond length, which depends on the character of both X and RO. Thus, the apical bond distance changes as much as 0.238 Å, from 1.742 Å in **1c** to 1.980 Å in **4a**, whereas the changes in the equatorial bond lengths (P···X$_{equatorial}$ and P···O$_{equatorial}$) are smaller than 0.070 Å for all the model compounds. Table 1 and Figure 1 also show that, for the same equatorial substitute, the P···OR$_{apical}$ bond length is shorter as the donor character of OR increases, while the donor character of the equatorial substitute X results in an increase of the P···OR$_{apical}$ bond length. Thus, for instance, for X = CH_3, the P···O$_{apical}$ bond distance changes from 1.814 Å in **1a** (apical substitute = OH) to 1.980 Å in **4a** (apical substitute = OCF_3), while for X = CH_3O, the P···O$_{apical}$ bond distance changes from 1.742 Å in **1c** (apical substitute = OH) to 1.883 Å in **4c** (apical substitute = OCF_3) (see Table 1 and Figure 1). In the case whether the equatorial substitute X is F, we have only found pentacoordinated compounds with apical substitutes HCOO (**3d**) and CF_3O (**4d**).

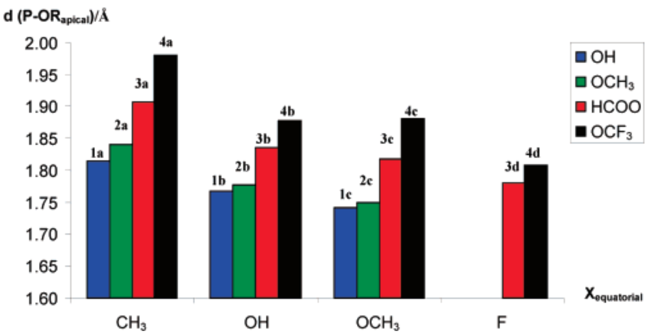
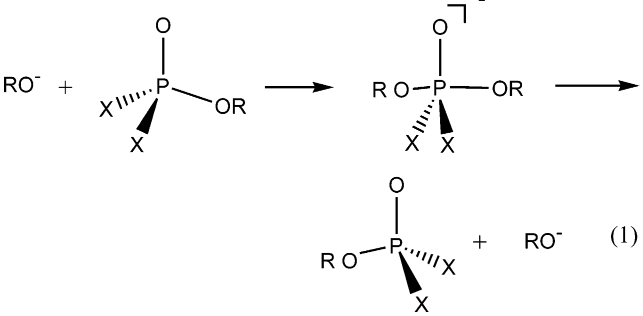


Figure 1. Diagram showing the dependence of the apical bond lengths in the pentacoordinated $POX_2(RO)_2$ model compounds on the nature of the apical (RO) and equatorial (X) substitutes on P.

These results indicate two important points, namely a different nature of the apical and equatorial bonds and a great importance of inductive effects on pentacoordinated phosphorus compounds. In order to get a deeper knowledge of the features of bonding at phosphorus, we have carried out a study of the bond properties according to the Atoms in Molecules (AIM) theory by Bader and the Natural Bond Orbital (NBO) theory by Weinhold. A detailed discussion of the AIM analysis is given in the Supporting Information along with the computed topological parameters at the bcp of the P···O$_{apical}$, the P···X$_{equatorial}$, and the P···O bonds collected in Table S4. In general, and regarding the apical bonds, the range of values for ρ_b and $\nabla^2\rho_b$ are characteristic of "closed shell" interactions, so that *the two apical bonds in these model systems can be classified as dative*. The large variability of the P···O apical bonds pointed out above, depending on the nature of the substitutes, is also typical of dative interactions.[71,72]

Additional information is provided by the NBO analysis. In Table 2 we have displayed the natural occupation at the P atom, the natural charge on P, and the stabilization energies ($\Delta E(2)$) associated with the charge-transfer interactions of the relevant donor−acceptor orbitals involving the apical bonds, that is, the bonding σ(P−O$_{apical}$), σ(P−X$_{equatorial}$), and σ(P−O) NBOs with the antibonding acceptor σ^*(P−O$_{apical}$) NBO. This stabilization energy has been computed with the second-order perturbation theory with the Fock matrix in the NBO analysis and the natural charges on phosphorus. The NBO analysis indicates that the natural occupation in the d shell is always less than or equal to 0.11 and consequently *excludes the participation of the d-orbital in the hybridization picture*. Thus, there is a formal sp^2 hybridization at P in all pentacoordinated model compounds. The d orbitals act as polarization functions in a similar way as pointed out by Reed and co-workers in a study on chemical bonding in hypervalent molecules[73] and in pentacoordinated silicon compounds bonded also by dative bonds.[72] This formal hybridization scheme is also compatible with the simple MO diagram based on a three-center four-electron (3c4e) model.[11,74] Another important point to be mentioned here refers to the topological features of the NBO orbitals linked to phosphorus. Those NBO orbitals designed as bonding orbitals of the type P···O or P···F in Table 2 are highly polarized toward the O or F atom, whereas those designed as antibonding orbitals have an almost exclusive contribution of phosphorus. Thus, the donor−acceptor interaction between these NBOs displayed in Table 2 represents quite well charge-transfer interactions. Moreover, this topological picture agrees very well with the dative description of the P···O bonds provided by the AIM analysis and discussed above. By the same way, the P···C bonds in compounds **1a**, **2a**, **3a**, and **4a** (with CH_3 as equatorial substitutes) have an almost equal contribution of phosphorus and carbon, according to the covalent character predicted by the AIM analysis (see above). The most important perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{PX-equatorial} \rightarrow \sigma^*_{PO-apical}$) are those having X = CH_3 (compounds **1a**, **2a**, **3a**, and **4a**) according to the well-known donor character of the methyl substitute and decreases according to the donor character of the equatorial substitutes (see above). Also very interesting are the perturbative donor−acceptor interactions between the two apical bonds ($\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$) and ($\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$) that involve charge transfer between the two apical bonds. Here it is also worth pointing out that these apical donor−acceptor interactions are symmetrical because the two apical groups are the same (see footnote b of Table 2). However, as will be shown below, when the two apical groups are different, the two apical donor−acceptor interactions are different, pointing out the competition of these two groups to form a dative bond to phosphorus and therefore having a direct influence on the corresponding P···O bonds.

Nucleophilic Substitution on the Model $POX_2(OR)_2$ Pentacoordinated Compounds. A very important point concerning the pentacoordinated $POX_2(OR)_2$ compounds discussed above refers to their relative stability. This has been studied in connection with the formation via a S_N2 reaction according to eq 1.



Please note that in this section, all reactions considered are symmetric as the entrance and leaving groups are identical. Each reaction described by eq 1 has been named according to the substitutes in the same way as has been done in the previous section to characterize the pentacoordinated phosphorus model compounds as displayed in Scheme 1. Thus, for instance, reaction **1a** means $RO^- = HO^-$ and X = CH_3, or reaction **4c** means RO = CF_3O^- and X = OCH_3; that is, each reaction has the same name that identifies the pentacoordinated intermediate. A schematic representation of the corresponding potential energy surfaces has been drawn in Figures 2 and 3, whereas the geometric parameters of the corresponding stationary points are collected in the Supporting Information. The energetic of these processes is contained in Table 3.

Figure 2a shows a schematic potential energy profile of reactions **1a**−**4a**, having the CH_3 group as equatorial substitute X. Each reaction begins with the formation of a prereactive hydrogen-bonded complex which occurs previous to the transition state and the formation of the pentacoordinate intermediate. Every prereactive complex has two hydrogen bonds, which occur between the oxygen of the anion (RO^-) and one of the hydrogen atoms of each equatorial methyl substitute. For reaction **3a** (red line, having HCOO as apical substitutes), the two hydrogen bonds in the prereactive complex are formed between each one of the oxygen atoms of the HCOO anion and one of the hydrogen atoms of each equatorial methyl substitute. The stability of these hydrogen-bonded complexes at 0 K is computed to vary among 24.3 and 16.6 kcal·mol^{-1} (for reactions **1a**−**4a**, see Table 3), and these energy values in gas phase are typical of hydrogen bond interactions involving an anion. After surmounting an energy barrier of the order of 5−6 kcal·mol^{-1}, the corresponding pentacoordinate intermediate is formed, and its stability, at 0 K, is computed to be among 33.1 and 15.9 kcal·mol^{-1}, relative to RO^- plus $POX_2(OR)$. The stability in these intermediates depends on the donor character of the apical substitutes. There is a large difference in the relative stability of the **1a** (blue line, apical substitute HO) and the stability of **4a** (black line, with apical substitute CF_3O), which amounts 16 kcal·mol^{-1}, so that *the compounds having the apical substitute with higher donor character are more stable*. This higher stability is associated with shorter apical bond lengths as discussed in the previous section for compounds **1a**, **2a**, **3a**, and **4a**.

In the case of the two equatorial (Figure 2b) substitutes X is the F atom, and we have only considered the reaction with $RO^- = HCOO^-$ (**3d**) and $RO^- = CF_3O^-$ (**4d**), since these are the only ones in which the F substitutes remain in the equatorial position as pointed out in the previous section. Both reactions occur by direct formation of a pentacoordinated phosphorus compound, whose stability at 0 K has been computed to be 32.1 and 20.7 kcal·mol^{-1}, for **3d** (red line) and **4d** (black line), respectively, according also to the higher donor character of the apical substitute in **3d** (see also Table 3). The processes are similar to those described recently by van Bochove and co-workers[24] in a recent study on nucleophilic substitution at phosphorus having fluorine atoms as equatorial substitute.

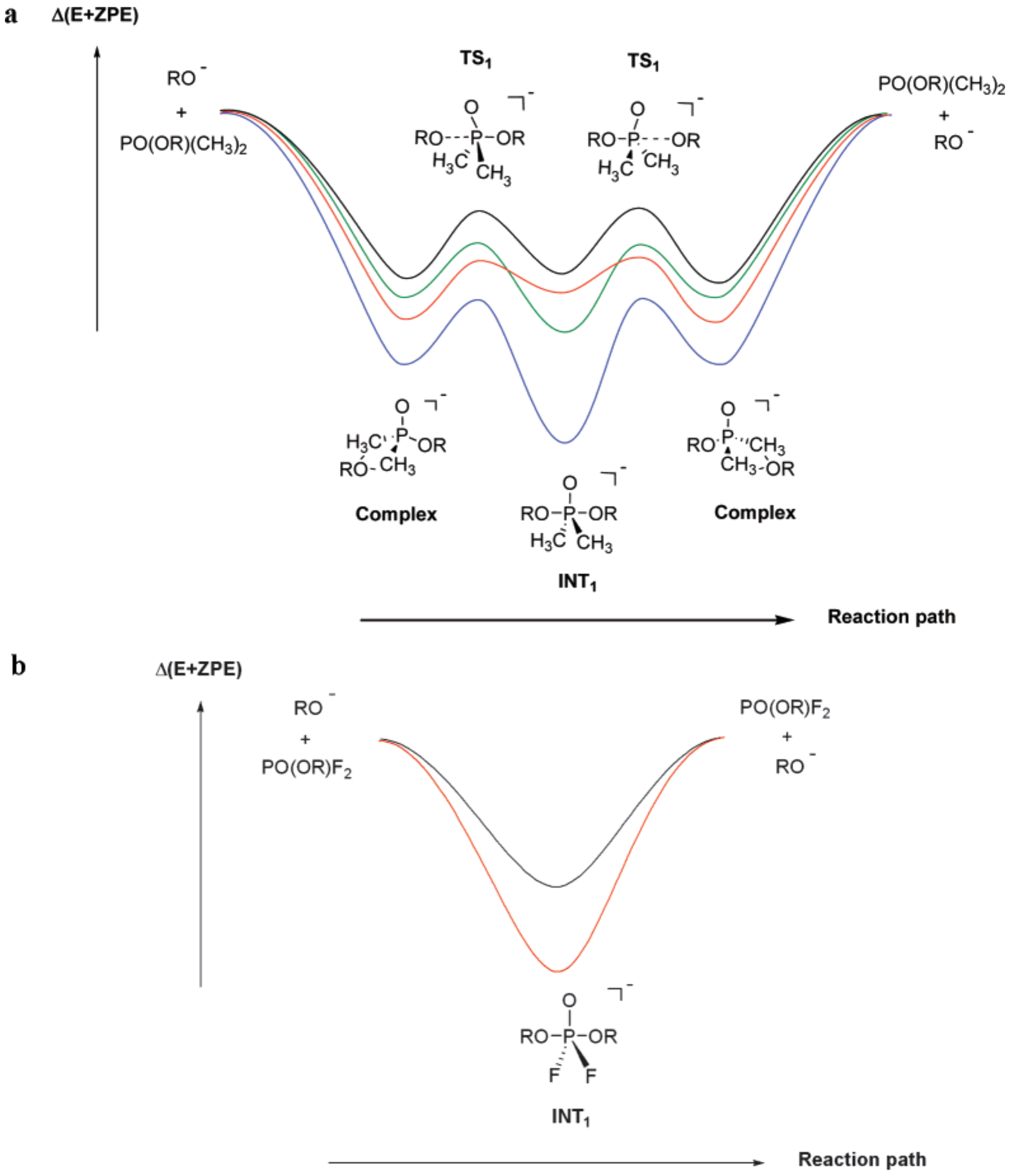


Figure 2. Schematic potential energy diagram for the nucleophilic substitution reactions: In (a), $RO^- + PO(OR)(CH_3)_2 \rightarrow PO(OR)(CH_3)_2 + RO^-$: (RO = HO, blue line; CH_3O, green line; HCOO, red line; and CF_3O, black line). In (b), $RO^- + PO(OR)(F)_2 \rightarrow PO(OR)(F)_2 + RO^-$: (RO = HCOO, red line; CF_3O, black line). The relative energies are computed at the *m*PW1PW91/6-311+G(3df,2p)//*m*PW1PW91/6-31+G(d) level of theory.

For the equatorial substitute X = OH, namely reactions **1b**−**4b**, the precursors of the corresponding pentacoordinated phosphorus compound are not RO^- and $PO(OH)_2OR$ as described in eq 1, but its respective conjugate acid (ROH) and basis ($PO(OH)(O)OR^-$), which occurs because $PO(OH)_2OR$ is a stronger acid than H_2O, CH_3OH, HCOOH, and CF_3OH, respectively (among 13.3 and 64.6 kcal·mol^{-1}, see reactions **1b**−**4b** in Table 3). Therefore, these model reactions involve a proton transfer linked to the formation of a pentacoordinated phosphorus compound, in a similar way as many reactions of biological interest. The schematic reaction profiles are depicted in Figure 3a, which shows that the reaction begins with the formation of a hydrogen-bonded complex which occurs previous to the formation of the pentacoordinated intermediate. This is a concerted process where the proton transfer from ROH to $PO(OH)(O)OR^-$ takes place simultaneously to the addition of the RO group to phosphorus. The results displayed in Table 3 and Figure 3a show that the computed stability of the prereactive hydrogen-bonded complexes ranges among 12.7 and 27.2 kcal·mol^{-1} and that the barrier that has to be overcome to form the pentacoordinated intermediate ranges among 37.1 and 27.5 kcal·mol^{-1} for **1b**−**4b**, respectively. Table 3 and Figure 3a show that all pentacoordinated intermediates lie energetically above the reactants (among 17 and −0.4 kcal·mol^{-1}). This reaction mechanism and the corresponding energetic profile is comparable to that of the dimethylphosphate hydrolysis and the ethylene phosphate hydrolysis reported recently.[26,28]

The last model reactions we have considered are those having X = CH_3O as equatorial substitutes and correspond to reactions **1c**−**4c**. A look at the schematic energy profile in Figure 3b shows that, for **1c**, **2c**, and **3c** (blue, green, and red lines respectively), the reaction has a 5-fold well. As

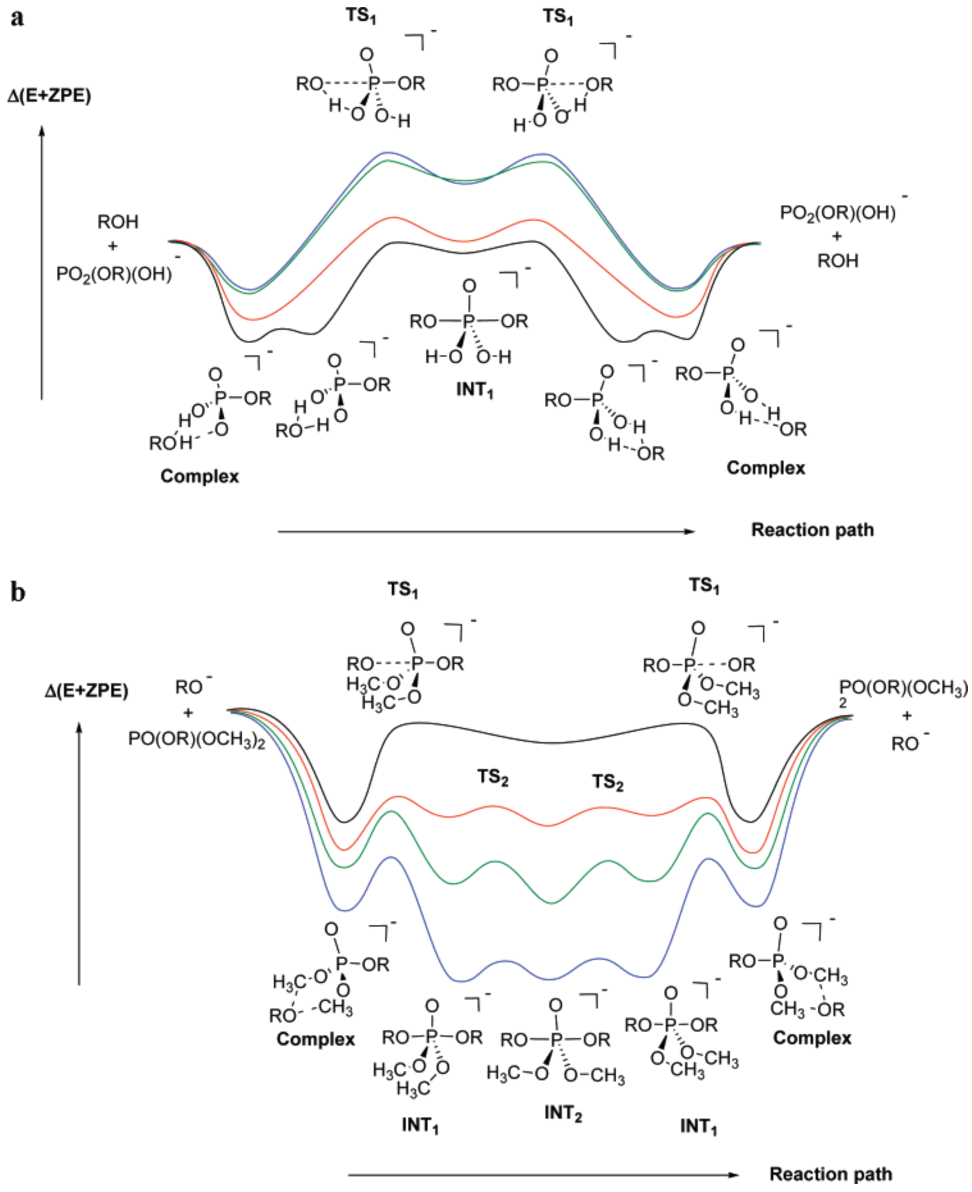


Figure 3. Schematic potential energy profiles for the nucleophilic substitution reactions: $RO^- + PO(OR)(X)_2 \rightarrow PO(OR)(X)_2 + RO^-$: ($RO = HO$, blue line; CH_3O, green line; HCOO, red line; and CF_3O, black line) in (a) X = (HO) and in (b) X = (CH_3O). The relative energies are computed at the *m*PW1PW91/6-311+G(3df,2p)//*m*PW1PW91/6-31+G(d) level of theory.

before, the reactions begin with the formation of a pentacoordinated hydrogen-bonded complex (first minimum), while the second, third, and fourth minima correspond to the pentacoordinated intermediates with the OCH_3 equatorial substitutes having different orientations, namely parallel to the side of the reactants (**INT1**), opposite (**INT2**), and parallel to the side of the products (**INT1**). The occurrence of similar multiple transition states separating the pentacoordinated species from the precursor complexes has been reported recently by van Bochove and co-workers,[24] who addressed this phenomenon to the increased steric bulk. For **4c** (black line) only one pentacoordinated intermediate has been found (**INT2**), being that the CH_3O equatorial substitutes are oppositely oriented. A more detailed discussion on these different conformers of the pentacoordinated phosphorus compounds will be given in the next section, and, for the aim of this section, it is only worth remarking here that the stability of the two pentacoordinated conformers differ only at most by 3 kcal•mol^{-1} (see Table 3). The results displayed in Table 3 reveal that the prereactive hydrogen-bonded complexes are computed to be among 26 and 15 kcal•mol^{-1} more stable than the reactants, and the formation of the pentacoordinated intermediate requires to surmount an energy barrier of among 7 and 12 kcal•mol^{-1}. Moreover, the stability of the pentacoordinated phosphorus intermediates follows the same trends as discussed above, namely that the intermediates having apical substitutes with higher donor character are more stable. That is 35.7 kcal•mol^{-1} for **1c** (apical substitute HO); 25.4 kcal•mol^{-1} for **2c** (apical substitute CH_3O); 15.2 kcal•mol^{-1} for **3c** (apical substitute HCOO); and 3.5 kcal•mol^{-1} for **4c** (apical substitute CF_3O). In addition, it is also worth mentioning that, as shown in Table 3, in the case of **1c** and **2c** (having apical substitutes with a large donor character) the pentacoordinated phosphorus intermediates are considerably more stable than the

Table 3. Relative Energies ($\Delta(E+\mathrm{ZPE})$ in kcal•mol^{-1}) Computed at the *m*PW1PW91/6-311+G(3df,2p)// *m*PW1PW91/6-31+G(d) Level of Theory for the Nucleophilic Substitution Reactions **1a**−**4d**

reaction[a]	RO^- + $POX_2(OR)$	ROH + $POX_2(OR)^-$	complex	TS1	INT1	TS2	INT2
1a	0.0		−24.3	−19.4	−33.1		
2a	0.0		−18.1	−13.3	−22.0		
3a	0.0		−20.4	−14.0	−17.9		
4a	0.0		−16.6	−9.9	−15.9		
1b	64.6	0.0	−12.7	24.4	17.2		
2b	52.4	0.0	−13.0	22.3	16.5		
3b	28.7	0.0	−18.8	5.7	1.3		
4b	13.3	0.0	−27.2	0.3	−0.4		
1c	0.0		−25.8	−18.9	−34.9	−32.7	−35.7
2c	0.0		−20.4	−13.2	−22.4	−19.9	−25.4
3c	0.0		−18.3	−11.5	−14.2	−12.7	−15.2
4c	0.0		−15.1	−2.2	−3.5		
3d	0.0				−32.1		
4d	0.0				−20.7		

[a] The following acronyms stand for the corresponding reactions: **1a** = HO^- + $OP(CH_3)_2(HO)$; **2a** = CH_3O^- + $OP(CH_3)_2(CH_3O)$; **3a** = $HCOO^-$ + $OP(CH_3)_2(HCOO)$; **4a** = CF_3O^- + $OP(CH_3)_2(CF_3O)$; **1b** = HO^- + $OP(OH)_2(HO)$; **2b** = CH_3O^- + $OP(OH)_2(CH_3O)$; **3b** = $HCOO^-$ + $OP(OH)_2(HCOO)$; **4b** = CF_3O^- + $OP(OH)_2(CF_3O)$; **1c** = HO^- + $OP(OCH_3)_2(HO)$; **2c** = CH_3O^- + $OP(OCH_3)_2(CH_3O)$; **3c** = $HCOO^-$ + $OP(OCH_3)_2(HCOO)$; **4c** = CF_3O^- + $OP(OCH_3)_2(CF_3O)$; **3d** = $HCOO^-$ + $OPF_2(HCOO)$; **4d** = CF_3O^- + $OPF_2(CF_3O)$.

prereactive hydrogen-bonded complexes, as opposed to what occurs for **3c** and **4c**.

Finally, it is also worth pointing out that the main reaction features described for these nucleophilic substitutions at phosphorus occur also in nucleophilic substitution reactions at silicon as reported by Bento and co-workers.[75]

Conformational Change in Equatorial Substitutes. The Polarization Effects. In the previous section we have pointed out that reactions **1c**−**3c** occur in several steps involving conformational changes in the orientation of the CH_3O equatorial substitutes. The corresponding energy barriers are smaller than 3.0 kcal•mol^{-1}, whereas the two conformers differ in energy at most by 3 kcal•mol^{-1} (see Table 3). Despite these small energetic differences in the two conformers, an analysis of its structures reveals significant differences with respect to the geometrical parameters concerning the apical substitutes. Therefore we have investigated the effect of the conformational changes (opposite and parallel orientation) on the equatorial substitutes in the model systems having HO and CH_3O as equatorial substitutes. In the case of the HO equatorial substitutes, only the model having HO as apical substitutes has both conformers stable (**1b** and **1b′**), whereas for the CH_3O equatorial substitutes the models with the HO, CH_3O, and HCOO apical substitutes have the two conformers stable (**1c** and **1c′**; **2c** and **2c′**; and **3c** and **3c′**; respectively). In Figure 4 we have displayed the most significant geometrical parameters of these conformers.

As pointed out in a previous section, the **1b** model has the two apical P•••O(H) bond lengths equal to 1.768 Å (see Table 1). However a conformational change in the equatorial substitute leading to a parallel orientation (model **1b′**) produces an important change in the two apical P•••O(H) bond lengths (1.695 and 1.910 Å, respectively); that is, the P•••O apical bond length opposite to the orientation of the two equatorial OH substitutes is reduced by 0.073 Å and the other P•••O apical bond length is enlarged by 0.142 Å. On the other hand, the changes in the equatorial bond lengths are very small (see Figure 4). From an energetic point of view, both conformers are separated by only 0.87 kcal•mol^{-1} ($\Delta(E + \mathrm{ZPE})$ value), being that **1b′** is more stable than **1b**. Looking for the origin of these differences we have first considered the possible existence of intramolecular hydrogen bond interactions that could stabilize one of these two conformers, but the AIM analysis ruled out this fact. Moreover, the NBO analysis indicates that the parallel orientation of the equatorial substitutes (structure **1b′**) induces a differential polarization effect on P, which results in a change of the phosphorus ability to bear an electronic charge and affecting therefore the axial bond length. In other words, the polarization on P produces a greater or less repulsion with the axial group (depending on the side) originating a change on the corresponding equilibrium bond distance. This polarization effect is not produced in those compounds with opposite oriented equatorial substitutes (structure **1b**) because of a cancellation effect due to the opposite orientation. For **1b**, the NBO analysis has already been reported in a previous section (see Table 2), where it has been pointed out that charge transfer occurs symmetrically. The perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{\mathrm{PO-equatorial}} \rightarrow \sigma^*_{\mathrm{PO-apical}}$) are equal to 22.6 kcal•mol^{-1} (from each of the two $\sigma_{\mathrm{PO-equatorial}}$ to each of the two $\sigma^*_{\mathrm{PO-apical}}$), whereas perturbative donor−acceptor interactions between the two apical bonds ($\sigma_{\mathrm{P1O5}} \rightarrow \sigma^*_{\mathrm{P1O6}}$ and $\sigma_{\mathrm{P1O6}} \rightarrow \sigma^*_{\mathrm{P1O5}}$) are both equal to 29.8 kcal•mol^{-1}. In the case of the conformer **1b′**, the situation changes radically, and the perturbative donor−acceptor interactions are not symmetrical anymore. The inductive effects, reflected in the perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{\mathrm{P1O3}} \rightarrow \sigma^*_{\mathrm{P1O6}}$ and $\sigma_{\mathrm{P1O4}} \rightarrow \sigma^*_{\mathrm{P1O6}}$), are 25.2 kcal•mol^{-1}, whereas ($\sigma_{\mathrm{P1O3}} \rightarrow \sigma^*_{\mathrm{P1O5}}$ and $\sigma_{\mathrm{P1O4}} \rightarrow \sigma^*_{\mathrm{P1O5}}$) are 20.5 kcal•mol^{-1}. That is, there is a greater charge transfer to the P1O6 side (σ^*_{P1O6} orbital) that affects the perturbative donor−acceptor interactions between the two apical bonds

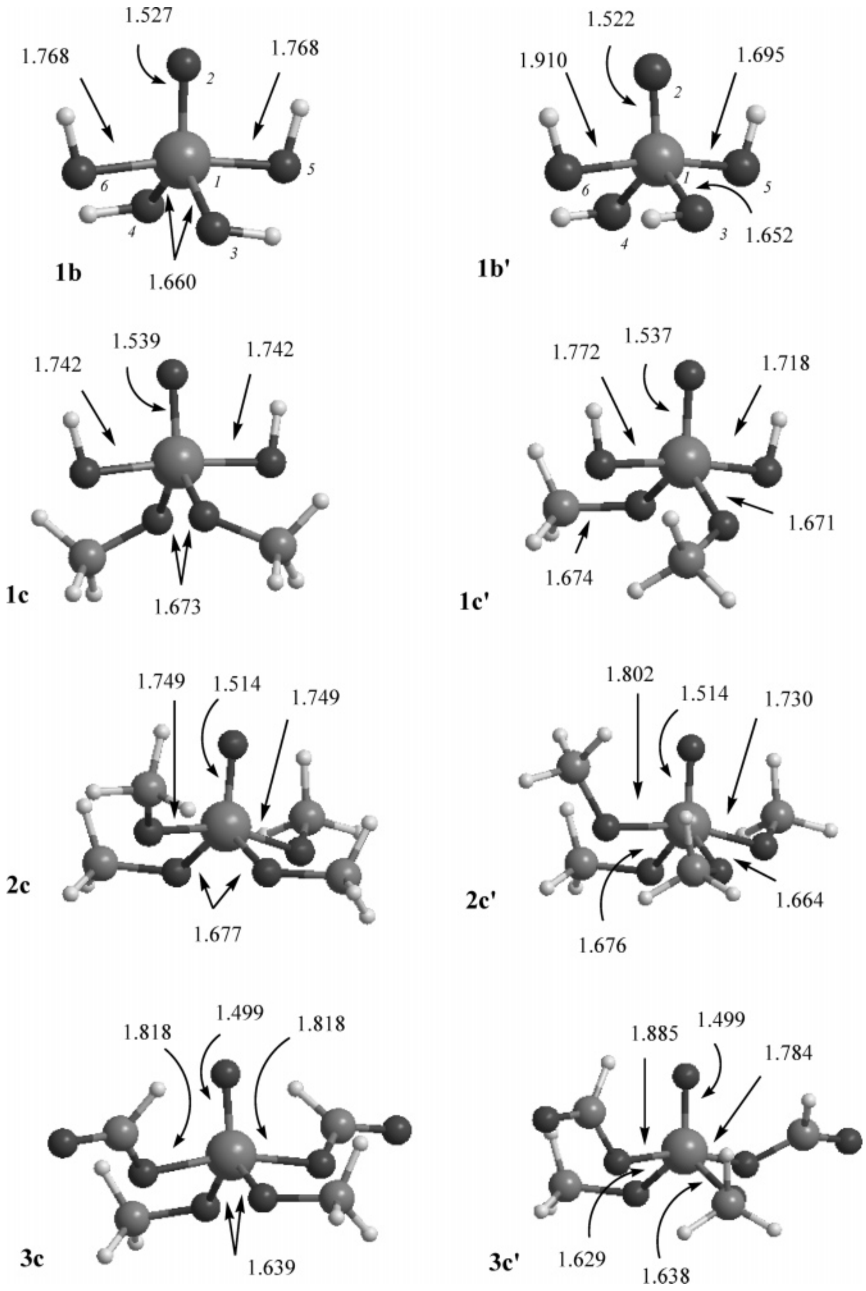


Figure 4. Selected geometrical parameters (in Å) for the optimized structures **1b**, **1b′**, **1c**, **1c′**, **2c**, **2c′**, **3c**, and **3c′**.

($\sigma_{P1O5} \rightarrow \sigma^*_{P1O6} = 32.1$ kcal·mol^{-1} and $\sigma_{P1O6} \rightarrow \sigma^*_{P1O5} = 27.7$ kcal·mol^{-1}), and, consequently, we can conclude that *the differential polarization effect originated by the conformational change induces a competition between the two equal apical substitutes in the pentacoordinated phosphorus compound*.

In order to visualize this induced polarization effect on P, we have considered the phosphoryl moiety derived from the two conformers **1b** and **1b′**, that is, we have deleted in both conformers the two apical substitutes. In the two resulting $PO(OH)_2$ moieties (one with the two HO opposite oriented and the other with the two HO parallel oriented) we have computed the molecular electrostatic potential (MEP), and the corresponding results are plotted in Figure 5. The phosphoryl having the two HO substitutes oppositely oriented, that derived from **1b**, (Figure 5a) has a symmetric distribution of the MEP in both sides of the equatorial plane, but, the phosphoryl group having the two HO substitutes parallel oriented, that derived from **1b′**, does not. Figure 5b,c

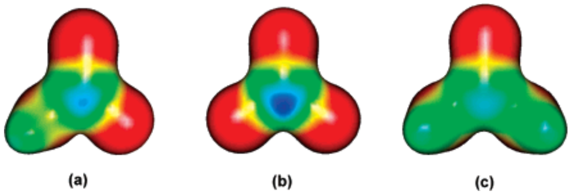


Figure 5. Molecular electrostatic potential representation of the $PO(OH)_2$ phosphoryl moiety of **1b** and **1b′** in a plane containing 98% of the electronic density: (a) one of the two symmetrical planes of **1b**; (b) opposite side of the equatorial hydrogens in **1b′**; and (c) side having the equatorial hydrogens in **1b′**.

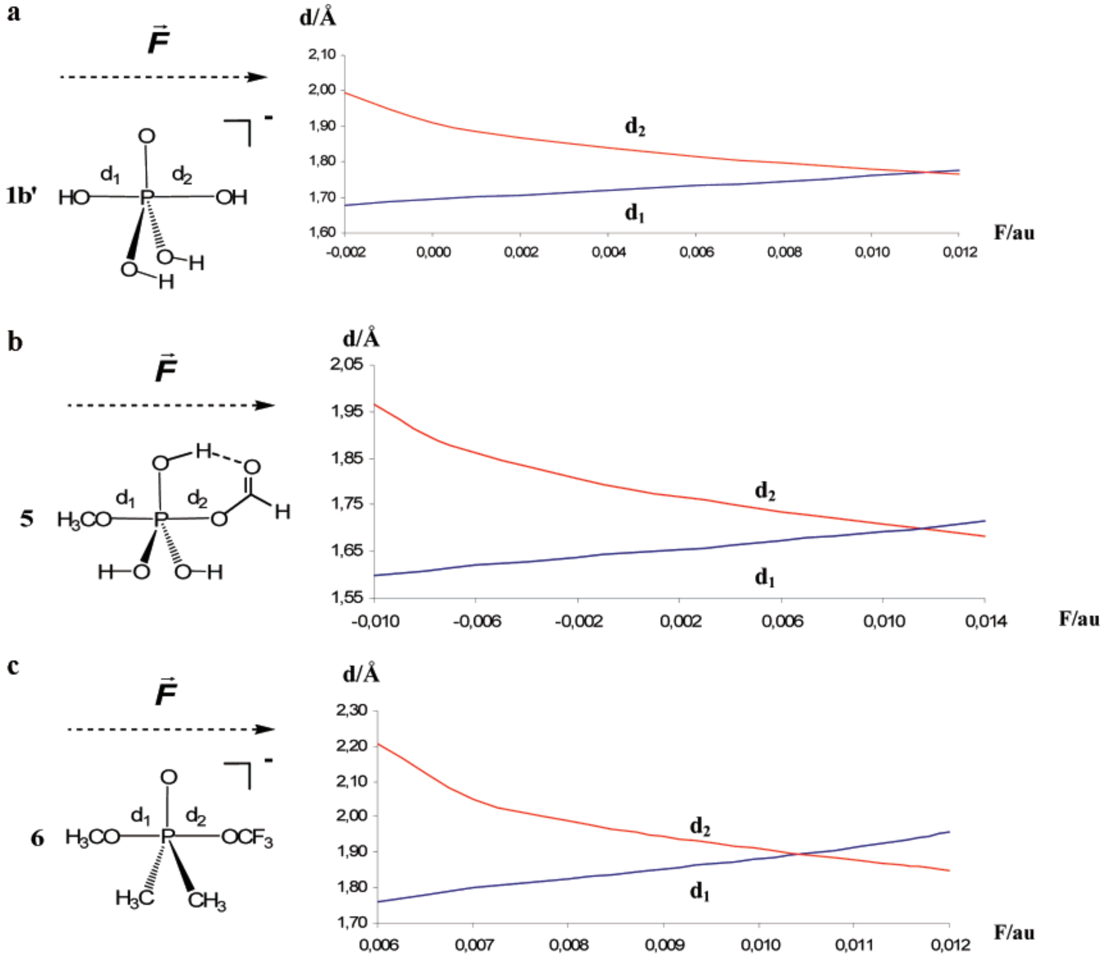


Figure 6. Dependence of the apical bond lengths on the intensity of an external electric field (*F*) for the pentacoordinated phosphorus compounds **1b′** (a), **5** (b), and **6** (c).

shows that it has a more positive charge density in the opposite side of the equatorial hydrogens. These results agree very much with the above discussion on **1b′**, where we have pointed out that a parallel orientation of the equatorial HO substitutes originates a large charge transfer to the side of the equatorial hydrogens (P1O6 bond in Figure 4).

A similar situation occurs with the compounds with equatorial substitutes CH_3O, structures **1c**−**3c**, and their corresponding conformers **1c′**−**3c′**. In a similar way as discussed above for **1b** and **1b′**, and as pointed out in the previous section, each pair of conformers differs at most by 3 kcal·mol^{-1}, being that the conformers **c** are more stable than the conformers **c′** (see Table 3). Figure 4 shows that compounds **1c**−**3c** have the CH_3O equatorial substitutes oppositely oriented and the two apical bond lengths equal (see also Table 1 and above), but a conformational change leading to the two equatorial substitutes parallel oriented (compounds **1c′**−**3c′**) produces, as just discussed for **1b** and **1b′**, a polarization effect on phosphorus that results in a significant change in the apical bond lengths. This is not so dramatic as for **1b** and **1b′**, because of the different electronegative character of the CH_3O equatorial substitutes, and the bond length changes induced amounts among 0.024 and 0.067 Å, depending on the apical substitutes (see Figure 4).

Effect of an External Electric Field. The high sensitivity to the polarization effects on the apical bonds, analyzed in the previous section, suggested to us to investigate the influence that an external electric field will produce on these kinds of bonds. To this end, we have performed a series of calculations on three pentacoordinated model systems and in two model reactions in order to analyze the effects of an external electric field on the geometries of the stationary point (minima) and on the reactivity. We have considered the effect of the external electric field in two different orientations, namely along a line in the plane defined by the phosphorus and the three equatorial substitutes and along the axis defined by the phosphorus and the apical substitutes. In the first case no substantial influence of the external electric field on the structures of the pentacoordinated models has been observed, but in the second case relevant effects have been found. Therefore, the results presented in this section correspond to the external electric field having the direction of the apical axis only. Putting the apical axis in the X direction and the origin of the coordinates at phosphorus, the external electric field follows the positive values of the X axis, while negative values means that the field direction was reversed. The results are displayed in Figures 6 and 7.

Regarding the influence of the electric field in the bonding and structural features, the first example we have considered

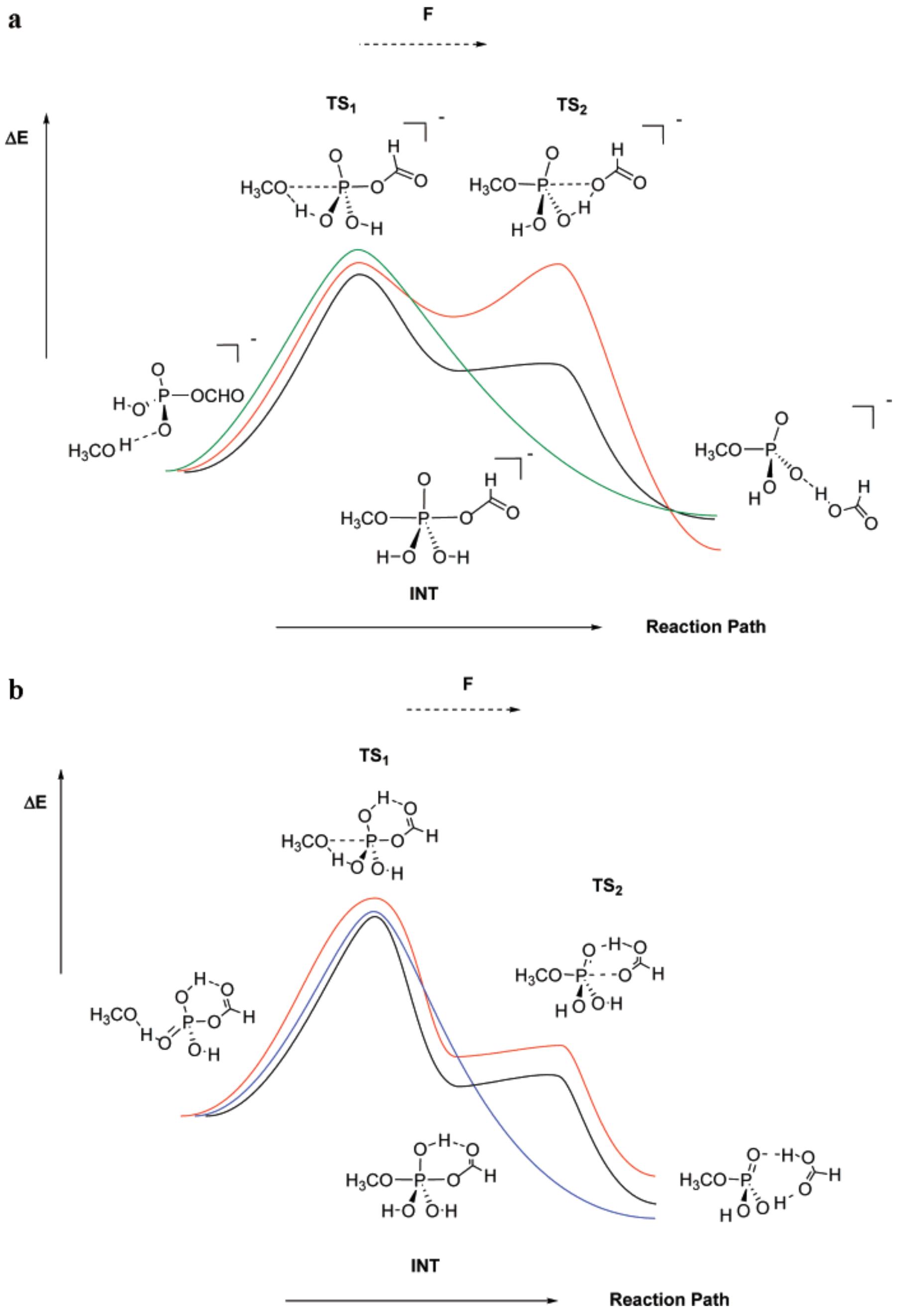


Figure 7. Schematic potential energy profiles computed under an external electric field at different intensities. (a) Corresponds to reaction 2 with $F = 0.0000$ au (black line); $F = 0.0060$ au (red line); and $F = -0.0020$ au (green line). (b) Corresponds to reaction 3 with $F = 0.0000$ au (black line); $F = 0.0060$ au (red line); and $F = -0.0060$ au (blue line).

is the model **1b′** ($PO(OH)_2(OH)_2$) discussed in the previous section and having the two equatorial OH substitutes parallel oriented (see Figure 4). We have pointed out that, in absence of external electric field, both apical P···O(H) bond lengths are different, 1.695 and 1.910 Å, respectively, for d_1 and d_2, but an external electric field produces important changes in these apical bond lengths. Figure 6a shows these changes as a function of the intensity of the external electric field. As the strength of F increases, d_1 is enlarged and d_2 is shortened so that applying an electric field of $F = 0.0111$ au both apical distances are equal, with a value of 1.769 Å. Figure 6a also shows that upon reversing the direction of the field, the opposite effect is observed, that is the d_2 is enlarged while d_1 is shortened, and with electric field with $F < -0.0030$ au, this pentacoordinated model is not stable anymore and dissociates in a process that involves a proton-transfer producing $H_2O + H_2PO_4^-$, as occurs in the process **1b** discussed in a previous section.

The second model we have considered under the effects of the electric field is $P(CH_3O)(HCOO)(HO)_3$ (compound **5**, see Figure 6b). This model is neutral, having three HO substitutes in an equatorial position, while the apical substitutes are CH_3O and HCOO. In the absence of an external electric field the two PO bond lengths are different (1.646 Å for P···OCH_3 and 1.785 Å for P···OCHO) as expected because, as pointed out above, the CH_3O apical substitute has a higher donor character. However, Figure 6b shows that the electric field produces a shortening of the P···O(CHO) bond length and a lengthening of the P···O(CH_3) bond distance so that with an external electric field of $F = 0.0107$ au the two apical P···O bond distances become equal to 1.699 Å. Figure 6b also shows that with fields with $F > 0.0107$ au P···OCH_3 becomes larger than P···OCHO, which means that the electric field changes the relative strength of the two apical bonds.

The third model we have considered is $PO(CH_3O)(OCF_3)(CH_3)_2$ having the two CH_3 as equatorial substituents and OCH_3 and OCF_3 as apical substituents (compound **6**, Figure 6c). This model has been chosen because in the absence of an external electric field, the pentacoordinated phosphorus compound is not stable and dissociates into $PO(CH_3O)(CH_3)_2$ and CF_3O^-. However, under a field of $F > 0.0060$ au, this pentacoordinated model is stable, being that the P···O(CH_3) bond length is shorter than the P···O(CF_3) until $F = 0.0105$ au, where both apical P···O bond distances become equal to 1.898 Å (see Figure 6c). When F increases beyond 0.0105 the P···O(CH_3) bond distance becomes larger than the P···O(CF_3) bond length, inverting thus the relative strength of the two apical bonds.

These three examples point out *a net influence of an external electric field on the bonding competence of the two apical dative bonds on phosphorus*.

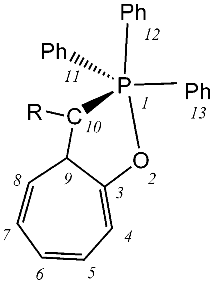
With regard to the study of the effect of F on the reactivity we have considered the two following nucleophilic substitutions:

$$CH_3OH + POO(OH)(OCHO)^- \rightarrow POO(OH)(CH_3O)^- + HCOOH \quad (2)$$

$$CH_3OH + PO(OH)_2(OCHO) \rightarrow PO(OH)_2(CH_3O) + HCOOH \quad (3)$$

These two reactions differ in the fact that in the second one we have added a proton in order to have a neutral reaction. The results are displayed in Figure 7. In both cases the reaction begins with the formation of a prereactive hydrogen-bonded complex, whereas in the exit channel a hydrogen-bonded complex is also formed before the release of the products. As we are mainly interested in what concerns the pentacoordination at phosphorus, we will consider these hydrogen-bonded complexes as reactive products of reactions 2 and 3. Moreover, as for reactions **1b**−**4b** discussed above, these reactions involve, in the entry and exit channels, a proton transfer which is linked to the formation (breaking) of the pentacoordination at phosphorus. For reaction 2, Figure 7a shows that in the absence of an external electric field (black line), the pentacoordinated phosphorus intermediate **7** is computed to be 17.2 kcal·mol^{-1} higher in energy than the prereactive complex. Its formation (via **TS1**) requires the surmounting of an energy barrier of 34.3 kcal·mol^{-1}, whereas the energy barrier for the exit channel (**TS2**) is only 1.3 kcal·mol^{-1}, that is, **TS1** is clearly the limiting step of the reaction. Figure 7a shows also that the reaction profile is significantly altered under an external electric field. Thus, with a $F = 0.0060$ au (red line), the pentacoordinated intermediate **7** is destabilized by about 9 kcal·mol^{-1}, and, more interestingly, the computed energy barrier for the exit channel (**TS2**) is the same as that of the back reaction (**TS1**) to the reactants. On the other side, with an external field of $F = -0.0020$ au (green line), the pentacoordinated phosphorus intermediate is not stable anymore, and the reaction occurs in a single step. A similar behavior is observed for the neutral reaction 3 (Figure 7b). In the absence of an external electric field, the reaction occurs through the pentacoordinated intermediate **6** (black line) and is slightly destabilized when a $F = 0.0060$ au is applied (red line). However, with an $F = -0.0060$ au (blue line) the pentacoordinated intermediate is not stable anymore, and the reaction occurs in a single step. These two examples point out that *the external electric field affects the stability of pentacoordinated phosphorus compounds and it may also affect the reactivity of nucleophilic substitution at phosphorus*.

Scheme 2 [a]



[a] **8a:** R = CH_3; **8b:** R = H; **8c:** R = CN.

These results may be of relevance in biological reactions involving pentacoordinated phosphorus, where the electric field originated by the folded protein could influence the catalytic process. In fact, it has been pointed out very recently the role of the electric field in the active site of the aldose reductase[76] and how the electric field may control the selectivity in heme enzymes.[77]

Triphenylphosphonium Ylide Derivatives. In an attempt to get a deeper insight in the hypervalence at phosphorus we have extended our investigation to the study on the bonding features of three neutral triphenylphosphonium ylide derivatives (**8**, Scheme 2) having pentacoordination at phosphorus and for which crystallographic X-ray data are available.[78,79]

These compounds have a trigonal bipyramid structure and are interesting for the purposes of this investigation because a change in the substitute R (R = CH_3 (**8a**), H (**8b**), and CN (**8c**)), which is not directly bonded to phosphorus, results in large changes in the P···O bond distance (2.00 Å for **8a**; 2.21 Å for **8b**; and 2.36 Å for **8c** (X-ray data)). The X-ray data first suggested that **8a** and **8b** form the PO bond but **8c** does not. Further analysis of the crystallographic data, together with results from ^{31}P and ^{13}C NMR spectra, had lead **8a**, **8b**, and **8c** to be viewed as resonance hybrids of structures A, B, and C (Scheme 3).[79,80] The δ_P and δ_C NMR spectra have been also collected in Table 4. **8a** shows a large δ_P (among −16.0 and −22.1 ppm, see also Table 4) that suggested a large contribution of the P−O bonding of the resonance structure A (Scheme 3). On the other hand, **8c** has a δ_P among −2.8 and −9.0 and has been related to the resonance structures B and C.

Scheme 3

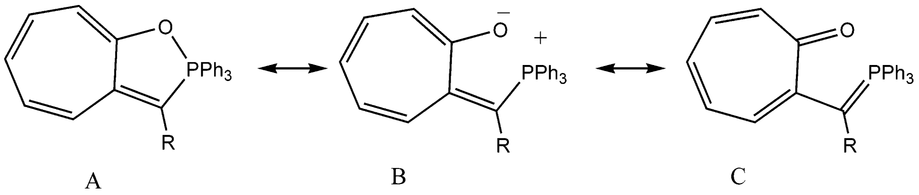


Table 4. Experimental and Computed ^{31}P and ^{13}C NMR Spectra (δ in ppm) for Compounds **8a**, **8b**, and **8c**[a]

	experimental values[69,70]					this work (gas phase)[b]		
	δ_P (CDCl$_3$)		δ_P (solid)	δ_C (CDCl$_3$)			δ_C	
compd	rt	−60 °C	rt	C3	C10	δ_P	C3	C10
8a	−17.9	−16.0	−22.1	170.1	87.0	−21.8	179.7	84.2
8b	−3.6	−1.8	−11.1	173.6	75.3	−16.2	180.2	77.8
8c	+7.8	+9.0	+2.8	177.5	49.4	−6.35	181.2	61.0

[a] Atom numbering is according to Scheme 2. [b] δ values are relative to H_3PO_4 for P and to TMS for C.

In the present work we have fully optimized and characterized as true minima the structures **8a**, **8b**, and **8c**, and their most significant geometrical parameters are displayed in Figure 8. It is gratifying to observe that the computed P···O bond distances compare well with the X-ray values for **8a** and **8b** (2.067 Å and 2.213 Å, respectively), whereas for **8c** our computed P···O bond length (2.239 Å) is 0.121 Å shorter than the X-ray value. Moreover, the remaining geometrical parameters compare also quite well with the experimental results. At this point, it should be taken into account that the calculated values should be compared with gas-phase values, while the X-ray data from the literature include packing effects that are shown to have an important role.[5,71] Besides the absolute values, the computed geometrical parameters follow the same trends with respect to the P···O bond lengths (**8a** < **8b** < **8c**). The bonding features have been analyzed, as above, employing the AIM and NBO methods, and the most significant results are displayed in the Supporting Information (Table S5). For each of the three triphenylphisphonium ylide considered (**8a**, **8b**, and **8c**), we have found a bcp between the phosphorus and oxygen atoms having the same topological features as those described in the previous sections for the P−O$_{\text{apical}}$ bonds in the model systems, that is, the values of the density and the Laplacian of the density are small and positive, indicating that there is a *PO bond, that can be classified as dative*. Moreover, and as above, the NBO analysis indicates that the phosphorus has a formal sp^2 hybridization scheme. On the other hand, the large differences in the P···O bond distances observed for the three compounds, and originated by the different substitutes R, can be mainly associated with the different ability to delocalize the π system through the seven-member ring. Thus, **8c** with R = CN has a certain amount of π character between C and N, which prevents, in part, the delocalization of the π system through the C9−C10 bond. This results in a shorter CO bond distance with less ability to transfer charge to phosphorus, and consequently the P···O bond distance is larger. The opposite case **8a**, with R = CH_3, implies a different delocalization of the π system through the seven-member ring resulting in a larger CO bond distance with more ability to transfer charge from oxygen to phosphorus and consequently with a smaller P···O bond length. In any case, these results show that small changes in the electronic features produce large changes in the P···O bonding.

For the sake of completeness we have also computed the ^{31}P and ^{13}C NMR spectra of **8a**, **8b**, and **8c**, and the results have been collected in Table 4 along with the experimental data. The computed NMR spectra correspond to gas-phase

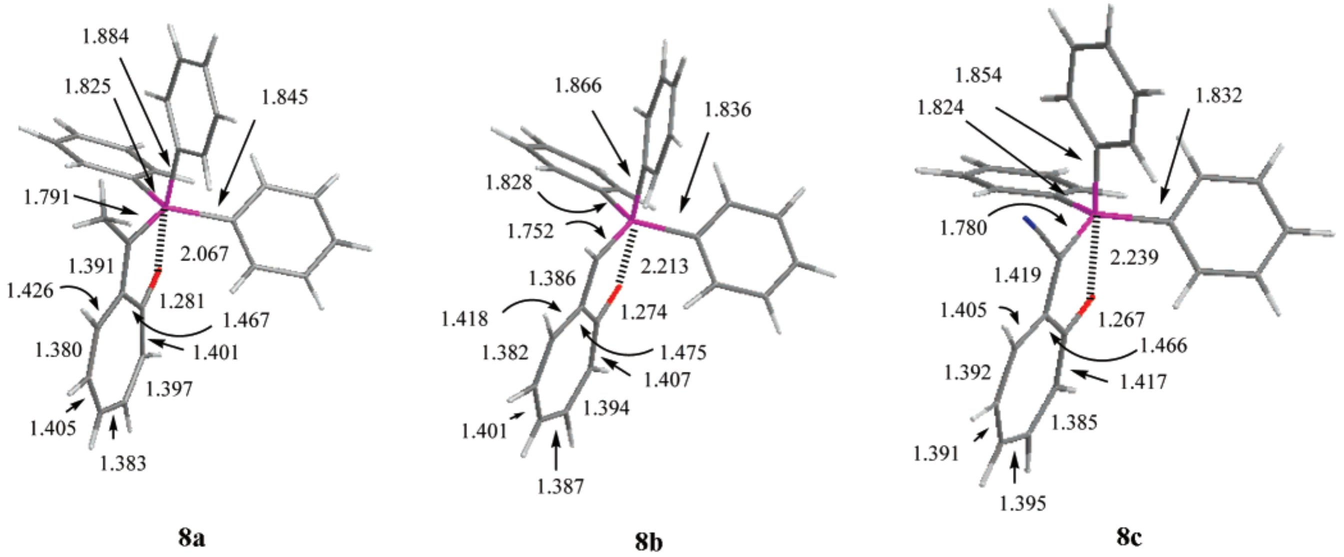


Figure 8. Selected geometrical parameters (in Å) for the optimized structures **8a**, **8b**, and **8c**. Atom numbering is according to Scheme 2.

optimized structures and show the same tendencies as the experimental values, that is larger δ_P for **8a** than for **8b** and for **8c** (−21.8, −16.2, and −6.35 ppm, respectively). Moreover, these results agree with the electronic features of the P←O dative bond. The stronger the P←O bond is, the higher the charge transfer associated with the dative bond and the shorter the corresponding bond length, which results in a higher shielding on P as reflected in the NMR spectra. It appears thus that the ^{31}P NMR spectra is a direct measure of the strength of dative bonding at phosphorus, and changes of its value in different media (as for instance in solid phase or in $CDCl_3$ for **8a**, **8b**, and **8c**, see refs 79 and 80 and also Table 4) would reflect differences in the bond length and strength. This also agrees with the linear correlation observed between δ_P and the X-ray P←O bond length as reported by Naya and Nitta.[79]

Conclusions

The results of the present investigation lead us to emphasize the following points: (1) All the pentacoordinated phosphorus compounds considered in this work have a trigonal bipyramid structure where the apical bonds show great variability. The topological and NBO analysis of the corresponding wave function indicates that these apical bonds can be classified as dative. These compounds are charge-transfer complexes, where the phosphorus has a formal sp^2 hybridization, which is compatible with the diagram based on a three-center four-electron (3c4e) model. (2) The features of the apical bonds depend strongly on the nature of the apical and equatorial substitutes. Compounds having apical substitutes with higher donor character are more stable and possess shorter apical bonds. On the other hand, the higher the donor character of the equatorial substitutes, the larger the apical bond length and the destabilization effect in pentacoordinated phosphorus compounds. (3) Polarization and electric field effects play an important role in the dative bonds of pentacoordinated phosphorus compounds, with consequences in both the geometry and the stability. These effects may change the competition between different apical substitutes, and they can even alter the reactivity of nucleophilic substitution at phosphorus. These effects may be of great relevance in enzymatic reactions, where the electric field originated by the folded protein could influence the catalytic process. (4) With regard to the three triphenylphosphonium ylide compounds considered (**8a**, **8b**, and **8c**), our results predict quite well the experimental (X-ray) geometrical data from the literature and show that in all cases there is a dative bond between the phosphorus and oxygen atoms, whose strength is correlated to the NMR displacement at P.

Acknowledgment. This research has been supported by the Generalitat de Catalunya (Grant 2005SGR00111). The calculations described in this work were carried out at the Centre de Supercomputació de Catalunya (CESCA) and at the Centro de Supercomputación de Galicia (CESGA), whose services are gratefully acknowledged. R.C. thanks also the Spanish Ramón y Cajal program.

Supporting Information Available: Cartesian coordinates of all structures reported in this paper and tables containing apical bond lengths for several model systems, optimized with different methods, activation and reaction energies for reaction **1b** obtained at different levels of theory, and AIM topological parameters for compounds **8**. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Skordalakes, E.; Dodson, G. G.; Green, D. S.; Goodwin, C. A.; Scully, M. F.; Hudson, H. R.; Kakkar, V. V.; Deadman, J. J. *J. Mol. Biol.* **2001**, *311*, 549−555.

(2) Lahiri, S. D.; Zhang, G. F.; Dunaway-Mariano, D.; Allen, K. N. *Science* **2003**, *299*, 2067−2071.

(3) Blackburn, G. M.; Williams, N. H.; Gamblin, S. J.; Smerdon, S. J. *Science* **2003**, *301*, 1184.

(4) Allen, K. N.; Dunaway-Mariano, D. *Science* **2003**, *301*.

(5) Holmes, R. R. *Acc. Chem. Res.* **2004**, *37*, 746−753.

(6) Leiros, I.; McSweeney, S.; Hough, E. *J. Mol. Biol.* **2004**, *339*, 805−820.

(7) Williams, N. H. *Biochim. Biophys. Acta* **2004**, *1697*, 279−287.

(8) Hengge, A. C.; Onyido, I. *Curr. Org. Chem.* **2005**, *9*, 61−74.

(9) Cleland, W. W.; Hengge, A. C. *Chem. Rev.* **2006**, *106*, 3252−3278.

(10) Wittinghofer, A. *Trends. Biochem. Sci.* **2006**, *31*, 20−23.

(11) Swamy, K. C.; Kumar, N. S. *Acc. Chem. Res.* **2006**, *39*, 324−333.

(12) Catrina, I.; O'Brien, P. J.; Purcell, J.; Nikolic-Hughes, I.; Zalatan, J. G.; Hengge, A. C.; Herschlag, D. *J. Am. Chem. Soc.* **2007**, *129*, 5760−5765.

(13) Mildvan, A. S. *Proteins* **1997**, *29*, 401−416.

(14) Allen, K. N.; Dunaway-Mariano, D. *Trends. Biochem. Sci.* **2004**, *29*, 495−503.

(15) Vedejs, E.; Marth, C. F. *J. Am. Chem. Soc.* **1988**, *110*, 3948−3958.

(16) Tremblay, L. W.; Zhang, G. F.; Dai, J. Y.; Dunaway-Mariano, D.; Allen, K. N. *J. Am. Chem. Soc.* **2005**, *127*, 5298−5299.

(17) Godfrey, S. M.; McAuliffe, C. A.; Pritchard, R. G.; Sheffield, J. M. *Chem. Commun.* **1998**, 921−922.

(18) Chandrasekaran, A.; Timosheva, N. V.; Day, R. O.; Holmes, R. R. *Inorg. Chem.* **2003**, *42*, 3285−3292.

(19) Hu, C. H.; Brinck, T. *J. Phys. Chem. A* **1999**, *103*, 5379−5386.

(20) Bianciotto, M.; Barthelat, J. C.; Vigroux, A. *J. Phys. Chem. A* **2002**, *106*, 6521−6526.

(21) Berente, I.; Beke, T.; Náray-Szabó, G. *Theor. Chem. Acc.* **2007**, *118*, 129−134.

(22) Wang, Y. N.; Topol, I. A.; Collins, J. R.; Burt, S. K. *J. Am. Chem. Soc.* **2003**, *125*, 13265−13273.

(23) Pepi, F.; Ricci, A.; Rosi, M.; Di Stefano, M. *Chem.-Eur. J.* **2004**, *10*, 5706−5716.

(24) Van Bochove, M. A.; Swart, M.; Bickelhaupt, F. M. *J. Am. Chem. Soc.* **2006**, *128*, 10738−10744.

(25) Lopez, X.; Schaefer, M.; Dejaegere, A.; Karplus, M. *J. Am. Chem. Soc.* **2002**, *124*, 5010−5018.

(26) Lopez, X.; York, D. M.; Dejaegere, A.; Karplus, M. *Int. J. Quantum Chem.* **2002**, *86*, 10−26.

(27) Lopez, X.; Dejaegere, A.; Leclerc, F.; York, D. M.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 11525−11539.

(28) Imhof, P.; Fischer, S.; Kramer, R.; Smith, J. C. *J. Mol. Struct. (THEOCHEM)* **2005**, *713*, 1−5.

(29) Grzyska, P. K.; Czyryca, P. G.; Golightly, J.; Small, K.; Larsen, P.; Hoff, R. H.; Hengge, A. C. *J. Org. Chem.* **2002**, *67*, 1214−1220.

(30) Chen, S. L.; Fang, W. H.; Himo, F. *J. Phys. Chem. B* **2007**, *111*, 1253−1255.

(31) Klahn, M.; Rosta, E.; Warshel, A. *J. Am. Chem. Soc.* **2006**, *128*, 15310−15323.

(32) Iche-Tarrat, N.; Ruiz-Lopez, M.; Barthelat, J. C.; Vigroux, A. *Chem.-Eur. J.* **2007**, *13*, 3617−3629.

(33) Cramer, C. J.; Gustafson, S. M. *J. Am. Chem. Soc.* **1993**, *115*, 9315−9316.

(34) Seckute, J.; Menke, J. L.; Emnett, R. J.; Patterson, E. V.; Cramer, C. J. *J. Org. Chem.* **2005**, *70*, 8649−8660.

(35) Uchimaru, T.; Tanabe, K.; Nishikawa, S.; Taira, K. *J. Am. Chem. Soc.* **1991**, *113*, 4351−4353.

(36) Yliniemela, A.; Uchimaru, T.; Tanabe, K.; Taira, K. *J. Am. Chem. Soc.* **1993**, *115*, 3032−3033.

(37) Tole, P.; Lim, C. M. *J. Phys. Chem.* **1993**, *97*, 6212−6219.

(38) Lim, C.; Tole, P. *J. Phys. Chem.* **1992**, *96*, 5217−5219.

(39) Lim, C.; Tole, P. *J. Am. Chem. Soc.* **1992**, *114*, 7245−7252.

(40) Chang, N. Y.; Lim, C. *J. Phys. Chem. A* **1997**, *101*, 8706−8713.

(41) Chang, N. Y.; Lim, C. *J. Am. Chem. Soc.* **1998**, *120*, 2156−2167.

(42) Dudev, T.; Lim, C. *J. Am. Chem. Soc.* **1998**, *120*, 4450−4458.

(43) Zhou, D. M.; Taira, K. *Chem. Rev.* **1998**, *98*, 991−1026.

(44) Taira, K.; Uchimaru, T.; Storer, J. W.; Yliniemela, A.; Uebayasi, M.; Tanabe, K. *J. Org. Chem.* **1993**, *58*, 3009−3017.

(45) Uchimaru, T.; Tsuzuki, S.; Storer, J. W.; Tanabe, K.; Taira, K. *J. Org. Chem.* **1994**, *59*, 1835−1843.

(46) Uchimaru, T.; Stec, W. J.; Tsuzuki, S.; Hirose, T.; Tanabe, K.; Taira, K. *Chem. Phys. Lett.* **1996**, *263*, 691−696.

(47) Range, K.; McGrath, M. J.; Lopez, X.; York, D. M. *J. Am. Chem. Soc.* **2004**, *126*, 1654−1665.

(48) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664−675.

(49) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

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

(51) Moeller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(52) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 281.

(53) Head-Gordon, M.; Head-Gordon, T. *Chem. Phys. Lett.* **1994**, *220*, 122.

(54) Ishida, K.; Morokuma, K.; Kormornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153.

(55) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(56) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

(57) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(58) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.

(59) Barlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1963.

(60) Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. *Int. J. Quantum Chem. XIV* **1978**, 545−560.

(61) Truhlar, D. G. *Chem. Phys. Lett* **1998**, *294*, 45−48.

(62) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 2921−2926.

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

(64) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. F. *C. J. Chem. Phys.* **1996**, *104*, 5497−5509.

(65) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251−8260.

(66) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R. J. A.; Montgomery, J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; 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.; Ocherski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, *Revision C.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(67) Shaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123−134.

(68) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899−926.

(69) Bader, R. F. W. *Atoms in Molecules. A Quantum theory*; Clarendon Press: Oxford, 1995; Vol. 22, pp 1−458

(70) Bader, R. F. W. *AIMPAC*. http://www.chemistry.mcmaster.ca/aimpac (accessed May 2002)

(71) Leopold, K. R.; Canagaratna, M.; Phillips, J. A. *Acc. Chem. Res.* **1997**, *30*, 57−64.

(72) Anglada, J. M.; Bo, C.; Bofill, J. M.; Crehuet, R.; Poblet, J. M. *Organometallics* **1999**, *18*, 5584−5593.

(73) Reed, A. E.; Schleyer, P. V. *J. Am. Chem. Soc.* **1990**, *112*, 1434−1445.

(74) Massey, A. G. In *Main Group Chemistry*; John Wiley and Sons: Chichester, England, 2000.

(75) Bento, A. P.; Bickelhaupt, F. M. *J. Org. Chem.* **2007**, *72*, 2201−2207.

(76) Suydam, I. T.; Snow, C. D.; Pande, V. S.; Boxer, S. G. *Science* **2006**, *313*, 200−204.

(77) Shaik, S.; de Visser, S. P.; Kumar, D. *J. Am. Chem. Soc.* **2004**, *126*, 11746−11749.

(78) Kawamoto, I.; Hata, T.; Kishida, Y.; Tamura, C. *Tetrahedron Lett.* **1971**, 2417-&.

(79) Naya, S.; Nitta, M. *J. Chem. Soc., Perkin Trans.* **2002**, *2*, 1017−1023.

(80) Nitta, M.; Naya, S. *J. Chem. Res.-S* **1998**, 522−523.

CT700220Z

JCTC Journal of Chemical Theory and Computation

Parallel Calculation of CCSD and CCSD(T) Analytic First and Second Derivatives

Michael E. Harding, Thorsten Metzroth, and Jürgen Gauss

Institut für Physikalische Chemie, Universität Mainz, Jakob-Welder-Weg 11, D-55099 Mainz, Germany

Alexander A. Auer*

Institut für Chemie, Technische Universität Chemnitz, Strasse der Nationen 62, D-09111 Chemnitz, Germany

Received June 25, 2007

Abstract: In this paper we present a parallel adaptation of a highly efficient coupled-cluster algorithm for calculating coupled-cluster singles and doubles (CCSD) and coupled-cluster singles and doubles augmented by a perturbative treatment of triple excitations (CCSD(T)) energies, gradients, and, for the first time, analytic second derivatives. A minimal-effort strategy is outlined that leads to an amplitude-replicated, communication-minimized implementation by parallelizing the time-determining steps for CCSD and CCSD(T). The resulting algorithm is aimed at affordable cluster architectures consisting of compute nodes with sufficient memory and local disk space and that are connected by standard communication networks like Gigabit Ethernet. While this scheme has disadvantages in the limit of very large numbers of compute nodes, it proves to be an efficient way of reducing the overall computational time for large-scale coupled-cluster calculations. In this way, CCSD(T) calculations of molecular properties such as vibrational frequencies or NMR-chemical shifts for systems with more than 1000 basis functions are feasible. A thorough analysis of the time-determining steps for CCSD and CCSD(T) energies, gradients, and second derivatives is carried out. Benchmark calculations are presented, proving that the parallelization of these steps is sufficient to obtain an efficient parallel scheme. This also includes the calculation of parallel CCSD energies and gradients using unrestricted (UHF) and restricted open-shell (ROHF) Hartree−Fock references, parallel UHF-CCSD(T) energies and gradients, parallel ROHF-CCSD(T) energies as well as parallel equation-of-motion CCSD energies and gradients for closed- and open-shell references. First applications to the calculation of the NMR chemical shifts of benzene using large basis sets and to the calculation of the equilibrium geometry of ferrocene as well as energy calculations with more than 1300 basis functions demonstrate the efficiency of the implementation.

1. Introduction

In electronic structure theory coupled-cluster (CC) methods have become a standard tool for high accuracy calculations.[1−5] With the exception of some difficult cases like multireference systems or cases of reference orbital instabilities, methods from the CC hierarchy represent robust black-box approaches providing increasing accuracy and a fast, systematic convergence to the full configuration-interaction (FCI) result. However, application of CC methods to larger chemical problems is limited by the rapidly increasing

* Corresponding author e-mail: alexander.auer@chemie.tu-chemnitz.de.

10.1021/ct700152c CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/16/2007

computational effort with growing number of electrons and basis functions.

A detailed analysis reveals that for CC methods like the coupled-cluster singles and doubles (CCSD)[6] and the coupled-cluster singles and double scheme augmented by a perturbative treatment of triple excitations (CCSD(T))[7] the limiting factor is CPU time and not storage requirements. If N is chosen as a measure of the system size, storage requirements for two-electron integrals, coupled-cluster amplitudes, and intermediates scale as N^4. In contrast to that, the operation count scales as N^6 for CCSD and N^7 for CCSD(T). Nowadays, efficient implementations allow calculations at the CCSD(T) level of theory with up to 800 basis functions.[8−11]

While for almost all methods numerous efficient parallel algorithms have been developed, the number of parallel implementations of CC methods has only increased in recent years.[12−19] Of note are the highly sophisticated algorithm for parallel calculation of CCSD(T) energies presented by Olson et al.[11] (within the program package GAMESS[18]) and an implementation of CCSD energies by Janowski et al.[20] (within the program package PQS[19]). In contrast to the algorithm presented in this work, the parallel implementation of CCSD(T) energies by Olson et al. is tailored to multiprocessor and/or multicore nodes connected by a dedicated communication network and based on the Distributed Data Interface (DDI/3).[11,21] And while Janowski et al. present CCSD energy calculations with more than 1500 basis functions on more than 30 compute nodes, their approach is based on the Array Files (AF)[20,22] scheme. We will focus on an ansatz which works without an additional layer of complexity provided by specialized libraries like DDI/3 or AF. The presented scheme is based on the message passing interface (MPI),[23] and all nonparallel steps run redundantly on every available processor at the same time.

To the best of our knowledge, however, no CC code capable of calculating general second-order molecular properties at the CC level using analytical derivatives has been adapted for parallel architectures. The main reason for this is that the mathematical structure of the CC equations makes an efficient fully parallel implementation or reimplementation demanding and time-consuming. In this paper we demonstrate an alternative approach, namely, the adaptation of an efficient serial algorithm to parallel environments.

The employed strategy is presented in a stepwise manner leading to an algorithm with parallelized routines for the time-determining steps in the CCSD and CCSD(T) energy, gradient, and analytical second-derivative calculations. We present benchmarks of large-scale CC applications using the Mainz−Austin−Budapest version of the ACES II program package[24] (ACES II MAB) modified in this way. A detailed investigation of the time-determining steps in CCSD and CCSD(T) calculations and the reduction of the overall execution time in the parallel algorithm is carried out.

2. Parallelization Strategy for CC Energies and Derivatives

A common approach for the parallelization of CC algorithms is to minimize storage requirements while aiming at constant (but in practice high) total communication by distributing parts of integrals, amplitudes, and intermediates and communicating the pieces as needed by other processors. While this approach guarantees a proper scaling of the algorithm in the limit of a large number of processors,[16,25] high-speed and expensive network connections are required. Furthermore, the structure of such an algorithm as well as the communication overhead, arising through dead times in which a node awaits data, may shift the crossover point with respect to efficient serial algorithms to a large numbers of nodes.

Following a different route, we apply a replicated storage scheme in order to minimize communication. Most of the quantities needed in the CC iterations are stored completely on every node in order to avoid communication of intermediate quantities. In contrast to the algorithm outlined by Olson et al.,[11] the work presented here is tailored to cluster architectures with moderate hardware specifications, assuming relatively slow interconnect structures like Gigabit Ethernet. Furthermore, it is assumed that memory is available to store the full set of T_1 and T_2 amplitudes locally on every node in fast memory and on hard disk. In this way, communication is minimized as only the CC amplitudes have to be communicated. The storage of the amplitudes rarely becomes a bottleneck: If one assumes a molecule with 20 occupied and 600 virtual spin orbitals, which would correspond to a basis set of better than quadruple-ζ quality, then the number of T_2 amplitudes, which scales as occ^2vrt^2, [26] would be of the order of a few hundred millions, which roughly corresponds to 1.5 GB of memory, if no symmetry is used.[27]

In the actual algorithm, parts of intermediates or integrals are contracted with the amplitudes on different nodes to give parts of the resulting quantity. In a final step, the amplitudes are updated and broadcast to all nodes. While allowing for distributed contractions during the CC iterations at minimal communication, this strategy has two drawbacks. Primarily, it does not allow for optimal scaling in the limit of a large number of processors as the distribution costs scale linearly with the size of the distributed entity and the number of processors. The exact scaling behavior for the communication depends on the employed communication hardware and the used algorithm. Furthermore, it does not reduce the storage requirements for the replicated quantities like the T_2 amplitudes or the molecular orbital (MO) integrals excluding the four-virtual index integrals. These are treated in partial atomic orbital (AO) algorithms which eliminate the need for a full transformation of the two-electron integrals and only require storage of the AO integrals. The needed MO integrals are calculated once in a semiparallel way and then are fully stored on each node (see subsections 2.2 and 2.3). It is straightforward to calculate and store AO integrals, which are usually the largest quantity in terms of disk space in modern CC algorithms, in a distributed manner. Together with the increased availability of large and cheap directly attached disk space the distributed storage of AO integrals makes it obsolete to recalculate or approximate these in every new step. At the same time, the efficiency of this algorithm is improving for increasing example size: The time-

Table 1. Timings for the Perturbative Triples Step in CCSD(T) Energies Relative to the Total Walltime of the CC Part

molecule	basis set	no. of electrons	no. of basis functions	% of (T) in CCSD(T)
H_2O	cc-pCVTZ	10	115	13
H_2O	cc-pCVQZ	10	144	13
H_2O	cc-pCV5Z	10	218	13
Cl_2	cc-pCVTZ	34	118	52
Cl_2	cc-pCVQZ	34	218	52
benzene	cc-pCVDZ	42	138	51
benzene	cc-pCVTZ	42	354	57
hexachlorobenzene	cc-pVDZ	168	192	60

determining steps can be more efficiently parallelized (see 3). However, calculations that are not feasible due to memory or disk space limitations (for the MO integrals) will also not be feasible when multiple processors are used.

For methods like CCSD or CCSD(T), communication costs associated with the replication of T_2-like quantities are usually at least 2 orders of magnitude smaller than the CPU time required for their computation. For CCSD(T), the scaling of CPU time is N^7 (occ^3vrt^4), while storage and communication costs grow as occ^2vrt^2 per compute node. Thus, the distribution of the time-determining steps to a number of processors in the way described above leads to a major reduction of overall walltime, especially when large examples are considered. A detailed discussion and examples for the different aspects of this parallelization strategy will be given in the next sections.

As the basic outline of the formalism used here and the common algorithms that are the starting point for our current work have been described in several publications,[7,28−32] we will not reiterate them but rather give details only for the steps modified in our approach.

In subsection 2.1 we will describe the parallelization of the CCSD(T) perturbative triples part for energies, gradients, and second derivatives starting from an implementation[28−30] that proves to be an ideal structure for the adaptation to parallel architectures.

In subsection 2.2 we carry out an analysis of the time-determining steps in CCSD energy, gradient, and second derivative calculations and describe the modification of the AO-based calculation of the leading term (the so-called particle−particle ladder term that includes contraction over two virtual indices).

In subsection 2.3 further issues for the optimization of the parallel code are described concerning the evaluation of two-electron integrals, the Hartee−Fock self-consistent-field (HF-SCF) procedure, and the integral transformation. For all test calculations reported in these sections correlation-consistent and correlation-consistent core-valence Dunning basis sets[33−35] have been used throughout.

Finally (section 3), we present applications of the new algorithm with a detailed investigation of the scaling of overall time with the number of processors.

2.1. Parallel Algorithm for the Perturbative Triples Contributions to CCSD(T) Energies, Gradients, and Second Derivatives.

The first step in the parallelization of the CCSD(T) scheme is to realize that almost all large CCSD(T) calculations are dominated by the calculation of the perturbative triples contribution. In Table 1 the timings for several standard serial CCSD(T) calculations are summarized. While the time-determining step for CCSD scales as occ^2vrt^4, the computational bottleneck of the perturbative triples correction scales as occ^3vrt^4. Thus, in comparison to CCSD the time spent for the (T)-correction more rapidly increases with the number of electrons, and this renders the computation of the perturbative triples correction the time-determining step in CCSD(T) calculations.

Our approach to parallelize the triples correction starts from the energy expressions[7]

$$E^{[4]} = \frac{1}{36}\sum_{ijk}\sum_{abc} t_{ijk}^{abc}\, D_{ijk}^{abc}\, t_{ijk}^{abc} \qquad (1)$$

$$E^{[5]} = \frac{1}{4}\sum_{ijk}\sum_{abc} \langle jk||bc\rangle t_i^a\, t_{ijk}^{abc} \qquad (2)$$

where $E^{[4]}$ and $E^{[5]}$ are energy contributions in fourth- and fifth-order perturbation theory, respectively. D_{ijk}^{abc} denotes the inverse orbital-energy denominator. As is the usual convention, i,j,k, ... denotes occupied and a,b,c, ... virtual spin orbitals. The perturbative-triple amplitudes t_{ijk}^{abc} are defined as

$$D_{ijk}^{abc}\, t_{ijk}^{abc} = P(k|ij)P(a|bc)\sum_{e} t_{ij}^{ae}\langle bc||ek\rangle - P(i|jk)P(c|ab)\sum_{m} t_{im}^{ab}\langle mc||jk\rangle \qquad (3)$$

with $P(x|yz)$ being the cyclic permutation operator ($P(x|yz)$ $f(x,y,z) = f(x,y,z) + f(y,z,x) + f(z,x,y)$), t_i^a and t_{ij}^{ab} the CCSD amplitudes, and $\langle bc||ek\rangle$ the antisymmetrized two-electron integrals. The basic scheme utilized in the ACES II MAB algorithm for the formation of the T_3 amplitudes is an outer loop over an index triple i,j,k of the t_{ijk}^{abc} amplitudes. For energy calculations, for example, blocks of a,b,c index triples are calculated within the loop one at a time and immediately used to form the $E^{(4)}$ and $E^{(5)}$ energy contributions. In this way, storage of the full triples amplitudes is circumvented, as has also been reported on many other occasions in the literature.[12,37,38]

If the T_1 and T_2 amplitudes and the corresponding integrals are fully or at least partially locally available on all nodes, each node can independently form a,b,c energy contributions. Only a single number per node, namely the summed up energy contributions, has to be communicated. In a final step, the energy contributions from the i,j,k blocks are summed up to give the total energy correction. In this way, the parallelization of the (T) energy contributions can be achieved in a straightforward fashion.

For CCSD(T) gradients, Watts et al.[38] describe an algorithm which following an idea of Lee and Rendell[39] avoids recomputation of amplitudes due to the use of perturbed canonical orbitals. Here, the outer loop runs again over the index triples i,j,k, and after the formation of an a,b,c block of T_3 amplitudes not only the energy increment but also the

Table 2. Timings for the Parallel Perturbative Triples Step in CCSD(T) Energy Calculations, Geometry Optimizations (One Iteration), and the Calculation of NMR Chemical Shifts as Analytical Second Derivatives for the Benzene Molecule[a]

	number of nodes				
	1	2	4	8	16
cc-pCVDZ (138)					
energy	75	40	19	9	5
geometry	251	126	63	31	16
NMR shieldings	2936	1452	727	363	192
cc-pCVTZ (342)					
energy	2163	1081	540	269	146
geometry	6594	3241	1619	809	426
NMR shieldings	80779	40285	20171	10527	5171
cc-pCVQZ (684)					
energy	29019	14514	7225	3592	1758
geometry	82171	40999	20425	10207	5489
cc-pCV5Z (1200)					
energy	238882	119469	59764	29895	15184

[a] Walltime in s. The number of basis functions is given in parentheses.

contributions to the unperturbed effective one-particle density matrices, the two-particle density matrices, and the contributions to the inhomogeneous terms of the Λ equations have to be calculated from the available T_3 block. In CCSD(T) second derivative calculations,[29,31] the same loop structure is used to construct the perturbed triple amplitues $\partial t_{ijk}^{abc}/\partial x$ and $\partial \tilde{t}_{ijk}^{abc}/\partial x$[40] and the corresponding contributions to the perturbed density matrices as well as to the perturbed Λ equations for one perturbation at a time. Such a strategy is only possible when using an asymmetric expression for the second derivatives.[29,31] This issue renders the asymmetric strategy the preferred choice over an alternative symmetric approach which requires the simultaneous availability of all perturbed amplitudes. However, while for energy calculations only one double-precision quantity needs to be communicated, for gradients and second derivatives the corresponding contributions to the two-particle density matrices must be exchanged and summed as well.

To illustrate the efficiency of this scheme, energy calculations, geometry optimizations, and calculations of NMR chemical shifts have been carried out for the benzene molecule using Dunning's correlation consistent core-valence basis sets.[41] The timings for the perturbative triples part of the algorithm are displayed in Table 2.[42]

The timings for the perturbative energy correction in the CCSD(T) algorithm scale almost perfectly up to 16 processors, even for the smallest basis set. It should be noted that the communication time required for the distribution of intermediate quantities calculated in the *a*,*b*,*c* loop of the perturbative triples is typically of the order of at most a few minutes, using Gigabit Ethernet interconnection. This is even the case for the largest examples and the largest numbers of nodes tested so far. In contrast to this, the time required for the parallel computation of the triples quantities themselves is usually of the order of hours for these examples.

In this way, using a simple scheme for the adaptation of the serial code to cluster architectures, the overall time for the most CPU-time intensive steps in CCSD(T) calculations can be scaled down efficiently. However, the required effort for the underlying CCSD calculation that precedes the calculation of the perturbative triples has not been discussed so far but now becomes the dominant step in the overall execution time. The next section focuses on this issue.

2.2. Analysis and Parallelization of CCSD Energy, Gradient, and Second-Derivative Calculations. From the previous section it is obvious that the straightforward parallelization of the (T) step in large-scale CCSD(T) calculations allows a significant reduction of the overall execution time up to a certain point. Increasing the number of nodes further, however, does not lead to an additional gain in execution time, if the effort for the underlying serial CCSD calculation exceeds the time for the parallel calculation of the perturbative triples. Thus, the next meaningful step in the parallelization of the CCSD(T) method is to identify and to parallelize bottlenecks in the CCSD algorithm. The time-determining steps in a CCSD energy calculation are the so-called particle−particle ladder terms that scale as $occ^2 vrt^4$ [32,43]

$$t_{ij}^{ab}\, D_{ij}^{ab} \leftarrow \frac{1}{2}\sum_{ef}\tau_{ij}^{ef}\langle ab||ef\rangle \tag{4}$$

where the intermediate

$$\tau_{ij}^{ab} = t_{ij}^{ab} + t_i^a\, t_j^b - t_i^b\, t_j^a \tag{5}$$

is used.

It should be noted that for CC energy and derivative calculations terms including $\langle ab||cd\rangle$ integrals or corresponding integral derivatives can, in general, be identified as the contributions with the highest scaling. For large basis sets the quartic dependence on the number of virtual indices will usually render this contraction expensive in terms of computational time.

One problem of the formulation in eq 4 is that the molecular-orbital integrals always represent a storage bottleneck, due to their lack of sparsity. As a consequence, the common practice in modern CC programs is an AO integral-driven algorithm in which the corresponding amplitudes are first partially transformed to the AO basis in an N^5 procedure

$$\tau_{ij}^{\mu\nu} = \sum_{ef} c_{\mu e} c_{\nu f} \tau_{ij}^{ef} \tag{6}$$

and then contracted with the AO integrals driven by the order in which integrals are retrieved from disk:

$$Z_{ij}^{\mu\nu} = \frac{1}{2}\sum_{\sigma\rho}\langle \mu\nu||\sigma\rho\rangle \tau_{ij}^{\sigma\rho} \tag{7}$$

Afterwards, the resulting intermediate will be back transformed and processed in the MO basis[44−50] as follows:

$$Z_{ij}^{ab} = \sum_{\mu\nu} c_{\mu a} c_{\nu b} Z_{ij}^{\mu\nu} \tag{8}$$

Olson et al.[11] give a detailed discussion on integral storage requirements and typical file size dimension.

Table 3. Timings per CCSD Iteration in Comparison to the Particle−Particle Ladder Term for the Benzene Molecule[a]

	number of nodes				
	1	2	4	8	16
	cc-pCVDZ (138)				
time per CCSD iteration	5.2	3.9	3.0	2.5	2.4
time for AO ladder term	3.4	2.0	1.1	0.6	0.5
	cc-pCVTZ (342)				
time per CCSD iteration	135	79	52	39	33
time for AO ladder term	113	57	31	16	9
	cc-pCVQZ (684)				
time per CCSD iteration	1902	1027	584	365	257
time for AO ladder term	1761	885	445	226	115

[a] Walltime in s. The number of basis functions is given in parentheses.

The use of partial AO algorithms has the advantage that only the more sparse AO integrals need to be stored at the expense of an additional transformation. While the operation count of the time-determining step scales as occ^2ao^4 in this scheme, in practice the reduced number of AO integrals also leads to a significantly reduced I/O and an overall saving of walltime. Thus, in almost all relevant cases the AO based algorithm outperforms the straightforward MO based scheme. Realizing that parallelizing this single contribution will lead to a major reduction of overall time in each CCSD iteration, we have chosen this AO based term as a starting point to improve our parallel CCSD(T) code.

The basic loop structure for which the power of multiple processors can be used effectively is an outer loop over batches of AO integrals that are read from disk and contracted with all matching T_2 amplitudes in the AO basis. After the contraction has been carried out, the resulting Z_{ij}^{ab} intermediate is communicated, and the CCSD iteration is continued.

It should be noted that the communication required after parts of the corresponding intermediate have been formed on all nodes scales at most as occ^2vrt^2 per compute node. It is expected that this step can be parallelized efficiently without running into communication bottlenecks.

Table 3 shows the timings of the CCSD iterations for the benzene molecule, where the AO ladder term has been parallelized in the described fashion.

As can be seen in the first column, the calculation of the AO-based particle−particle ladder term dominates the time for one iteration, even for the smallest examples by more than 60%. Furthermore, the parallelization of this term in batches of AO integrals results in an almost perfect reduction of the walltime for this contribution up to 16 processors and, thus, to a significant reduction of the overall time per iteration, especially for larger examples.

For the calculation of analytic gradients terms analogous to those in energy calculations appear in the CCSD Λ equations:[32]

$$\lambda_{ij}^{ab}\, D_{ij}^{ab} \leftarrow \frac{1}{2}\sum_{ef}\lambda_{ij}^{ef}\langle ef||ab\rangle \qquad (9)$$

Table 4. Timings for the Solution of the Lambda Equations and the Solution of the Perturbed Amplitude and Lambda Equations for the Benzene Molecule[a]

	number of nodes				
	1	2	4	8	16
	Lambda Equations				
cc-pCVDZ	65	48	40	35	34
cc-pCVTZ	1819	1147	828	653	603
cc-pCVQZ	25166	14808	9510	6887	5854
	Perturbed Amplitude and Lambda Equations				
cc-pCVDZ	607	459	390	345	331
cc-pCVTZ	18526	12190	9255	7972	6022

[a] Walltime in s.

Within the ACES II MAB program package this term is calculated using the same AO integral-driven scheme. Thus, the time-determining N^6 step in the gradient calculations can be parallelized in the same way as the corresponding term in the energy calculation.

For second derivative calculations, the contributions that have to be considered occur in the equations for the perturbed cluster and perturbed Λ amplitudes:

$$\frac{\partial t_{ij}^{ab}}{\partial\chi}\, D_{ij}^{ab} \leftarrow \frac{1}{2}\sum_{ef}\frac{\partial t_{ij}^{ef}}{\partial\chi}\langle ab||ef\rangle \qquad (10)$$

$$\frac{\partial\lambda_{ij}^{ab}}{\partial\chi}\, D_{ij}^{ab} \leftarrow \frac{1}{2}\sum_{ef}\frac{\partial\lambda_{ij}^{ef}}{\partial\chi}\langle ab||ef\rangle \qquad (11)$$

By parallelizing these contributions in the AO based scheme, the overall computational cost of the most time-consuming steps in the CCSD gradient and analytical second derivative calculations can be reduced as well. Due to the dominance of these steps compared to the overall time per CCSD iteration this simple strategy improves the overall CCSD time significantly if multiple processors are used.

Table 4 summarizes the timings for the corresponding modules that include the steps described above for calculations of NMR chemical shifts for the benzene molecule.

From the last columns of Tables 2 and 4 it becomes clear that for this example the overall time required for the CCSD- and CCSD(T)-derivative equations is of the same order as the time for the evaluation of the perturbative triples part, if 16 processors are used. Thus, the CCSD part of the calculation will still dominate the overall time for CCSD(T) derivative calculations if more than 16 processors are used. This is mainly due to contributions that have not been parallelized, which include terms of lower scaling, integral derivative transformations, etc.

So far only the particle−particle ladder terms for the CCSD, Λ, perturbed amplitudes and perturbed Λ equations are parallelized.

To further improve the algorithm, one could utilize parallel matrix multiplication routines for CCSD contributions that scale as occ^3vrt^3. In addition, for the special case of analytic second derivatives, one could also use a coarse-grained parallelization scheme[51] on top of the one suggested here.

Table 5. Timings for the Calculation of Two-Electron Integrals, the HF-SCF, and the Integral Transformation for the Benzene Molecule[a]

	number of nodes				
	1	2	4	8	16
	cc-pCVDZ (138)				
integral evaluation	8.5	4.6	2.2	1.1	0.7
HF-SCF	1.8	1.1	0.6	0.4	0.3
integral transformation	1.8	1.4	1.3	1.1	1.3
	cc-pCVTZ (342)				
integral evaluation	340	167	90	45	24
HF-SCF	57	30	16	9	5
integral transformation	61	37	29	23	21
	cc-pCVQZ (684)				
integral evaluation	19379	10012	4906	2518	1505
HF-SCF	1009	453	225	122	67
integral transformation	1096	601	399	267	233

[a] Walltime in s. The number of basis functions is given in parentheses.

Namely, one could distribute the different perturbations that are calculated independently, for example B_x, B_y, and B_z for NMR chemical shifts or geometric perturbations in the case of harmonic frequencies, to different processors. While this has not been carried out in the approach presented here, it is a straightforward addition to any code that could further help to improve the scaling of the algorithm with the number of processors.

However, the crossover point from which the preceding CCSD calculation will dominate over the CCSD(T) calculation time is pushed to a larger number of processors for larger examples. For larger molecules with medium-sized basis sets this crossover point should shift to 32 or even 64 processors, so that large-scale cluster architectures could be used to carry out calculations within days that would take months on a single processor.

2.3. Further Optimization Issues. The scheme for parallelizing the AO ladder terms described in subsection 2.2 requires only equally distributed integrals to be present on the different compute nodes. As a consequence, the evaluation of the integrals is carried out in parallel and in turn used in the parallel framework of the HF-SCF procedure and the integral transformation. Each node calculates and stores only a part of the integrals, and thus during the HF-SCF procedure only an incomplete Fock-matrix is built on each node, which is then exchanged between the compute nodes. The total Fock matrix is simply the sum of these incomplete matrices. The rest of the algorithm is unaltered. For the integral transformation all AO integrals are read in only once and communicated in the form of an intermediate array. After transformation of the MO integrals with two and three virtual indices each node stores all calculated MO integrals locally. Another non-negligible part of the derivative calculation is the evaluation of the integral derivatives which can be parallelized in an analogous manner. While this has not been done in the work presented here, it is the focus of future work, among other optimization issues.

The timings (in seconds) for parallel integral evaluation, HF-SCF, and integral transformation for the benzene molecule are shown in Table 5. This simple scheme results not only in a reduced storage requirement per node but also in a reduction of the overall time for the integral evaluation, transformation, and HF-SCF. It should be noted, that for some cases, even superlinear scaling of the evaluation of the two-electron integrals can be observed. This is due to automatic buffering schemes in the operating system that allow for a more efficient I/O if certain buffer sizes are reached and has also been reported by other groups.[52]

An important issue in parallel implementations is to avoid load balancing problems. In the work presented here, HF-SCF, the integral transformation, and any CCSD-like equations are automatically balanced by the local calculation and storage of equally sized amounts of two-electron integrals at the beginning of the calculation. The remaining steps in the calculation of the perturbative triples are balanced on average by the large number of these contributions. This applies for the calculation of energies, gradients, and any second-order properties. In practice, even multiprocessor systems do not show balancing problems since the load is kept equal on every processor. For the actual implementation we assume that dedicated nodes are available. However, load balancing problems will arise if heterogeneous resources are used or if compute nodes have different loads due to other calculations. This issue will have to be addressed in further developments of the current algorithm.

3. Results and Discussion

In this section we focus on the overall performance of the scheme presented here and the practicability for the usage on typical cluster architectures. The results of two applications are presented that outline typical problems in quantum chemistry for which high level ab initio methods are necessary but extremely time-consuming unless parallel implementations, like the one presented here, are applied.

Nowadays more than 300 GB of disk space and 8 GB of random access memory (RAM) are readily available even on single cluster nodes within medium sized computer clusters, so it is not foreseeable that memory or storage will present a bottleneck for larger calculations. As has been stressed before, only serial calculation times of the order of months or years will render large-scale CCSD(T) calculations infeasible. While a parallel implementation cannot combat the steep scaling of high-level CC methods, the power of parallel computer architectures can help to stretch the range of applicability far beyond what it has been in recent years.

In Table 6 the results of some representative benchmark calculations for energies and gradients are summarized. The timings for CCSD(T) energy calculations for benzene and cyclohexene and the timings of one step of the geometry optimization of the adamantyl cation ($C_{10}H_{15}^{+}$) are given.

For the high-symmetry case benzene in a hextuple-zeta basis set the serial energy calculation would take about 2 weeks and is reduced to less than a day using 16 processors. From Table 6 it also becomes obvious that the number of basis functions is not the only factor when considering the size of a system but also symmetry and the distribution of orbitals among the irreducible representations as well as the ratio of occupied to virtual orbitals.

Table 6. Overall Timings for the CCSD(T) Energy Calculations of Benzene and Cyclohexene and for One Step of the Geometry Optimization of the Adamantyl Cation

molecule	comput. symm	basis set	no. of basis functions	no. of nodes	execution time [h]
benzene[a]	D_{2h}	cc-pV5Z	876	16	4
benzene[a]	D_{2h}	cc-pV6Z	1386	16	21
cyclohexene[a,b]	C_2	aug-cc-pVQZ	940	16	40
adamantyl cation	C_s	cc-pVTZ	510	9	90

[a] Energies (frozen core) for benzene cc-pV5Z, cc-pV6Z, and cyclohexene are −231.8916163, −231.8987752, and −234.2797258 Hartree. [b] Carried out at the fc-MP2/cc-pVTZ geometry.

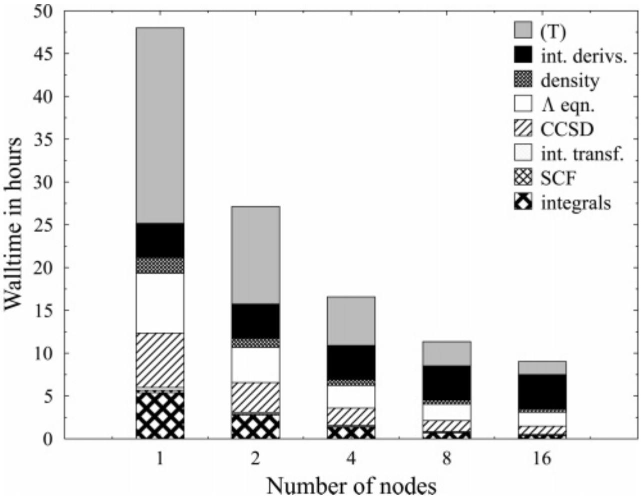


Figure 1. Composition of the overall walltime for one step in the geometry optimization of the benzene molecule at the CCSD(T)/cc-pCVQZ level of theory (684 basis functions).

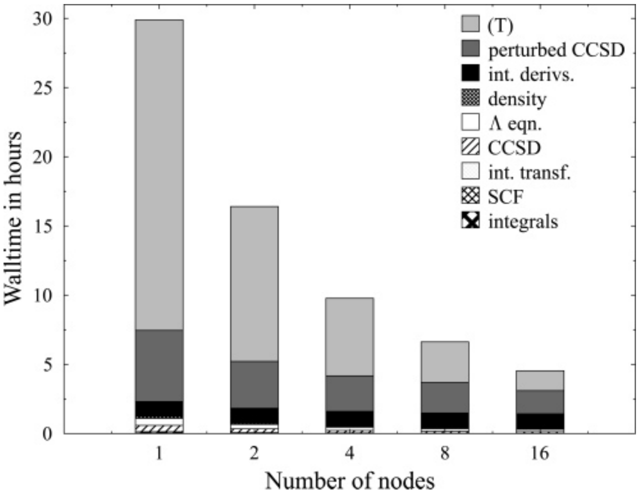


Figure 2. Composition of the overall walltime for the calculation of the NMR chemical shifts of the benzene molecule at the CCSD(T)/cc-pCVTZ level of theory (342 basis functions).

Figures 1 and 2 give a more detailed view on the different steps required for two smaller CCSD(T) calculations, namely the gradient for one step of a geometry optimization and for the calculation of NMR chemical shifts also for the benzene molecule.

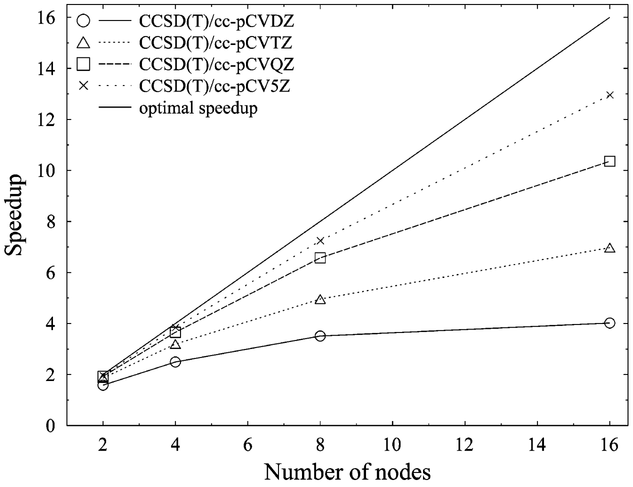


Figure 3. Parallel scaling of CCSD(T) energy calculations for the benzene molecule using different basis sets.

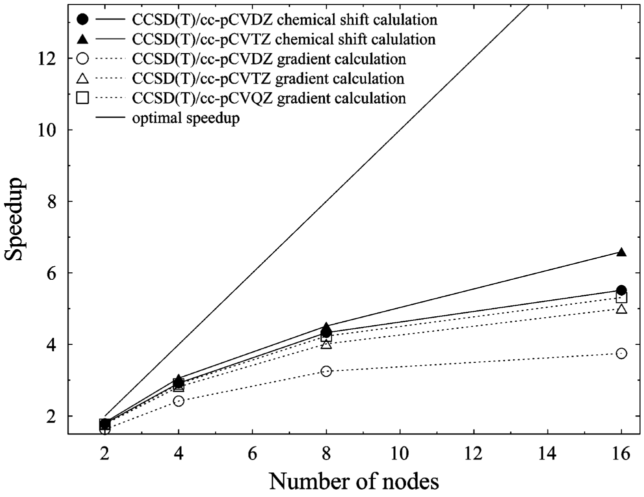


Figure 4. Parallel scaling of CCSD(T) first and second analytical energy derivative calculations for the benzene molecule using different basis sets.

For the gradient calculation, which takes about 47 h on a single processor, it can be seen that the calculation of the perturbative triples contribution takes only about half of the overall time of the optimization step. Using the current algorithm it is possible to scale down this contribution and also the calculation of the two-electron integrals, the CCSD, and Λ equations as well as the integral transformation and the contributions to the density matrices. Using 16 processors the overall time is reduced to less than 10 h. At this point steps dominate the overall time that have not been considered for parallelization in the algorithm, so that the usage of larger numbers of nodes would not yield significant further speedups.

The calculation of the NMR chemical shifts with a larger basis set shows a different profile. Here the perturbative triples contributions to the second derivatives clearly dominate compared to the other steps, such as the SCF or CCSD calculations. Using 16 processors the overall time of 30 h can be reduced to less than 5 h. After this point, the remaining steps in the perturbed CCSD equations that have not been parallelized dominate the overall calculation time.

Table 7. ^{13}C NMR Chemical Shifts for the Benzene Molecule Using Various Basis Sets[41] [a]

basis	no. of basis functions	absolute NMR shieldings CCSD(T)
cc-pCVDZ	138	82.07
cc-pCVTZ	342	66.61
tz2p[c]	198	68.42
qz2p[c]	228	64.95
pz3d2f[c]	474	63.22
vib corr[b]		−3.43
total		59.79
experiment		57.1[57]

[a] The experimental values have been taken from ref 57 using the absolute shifts of carbon monoxide ($\sigma_{T=300K}$=0.9 ± 0.9 ppm).[58] For a detailed description of the basis sets used and the scheme for the computation of the zero-point vibrational correction see ref 53. To avoid the gauge-origin problem in the calculation of NMR chemical shifts the gauge including atomic orbitals (GIAO)[59−61] approach has been used. [b] The vibrational correction is based on a perturbational approach.[53] The cubic force field was calculated at the MP2/cc-pVTZ and the NMR shieldings for the displacements at the MP2/qz2p level of theory. [c] The qz2p basis consists of a 11s7p2d/7s2p primitive set contracted to 6s4p2d/4s2p and the pz3d2f basis of a 13s8p3d2f/8s3p2d set contracted to 8s5p3d2f/5s3p2d.[62−65]

Figures 3 and 4 give detailed insight for the scaling of CCSD(T) energy and derivative calculations. As has been discussed in the previous sections, the scaling of total time with the number of nodes is improving for increasing system size as the importance of the parallelized, time-determining steps is even larger. Thus, this implementation will make applications feasible within weeks or even days that would take months or years to calculate, if 64 or 128 nodes were used.

The following examples give the proof of principle, that this simple scheme for adapting a serial implementation leads to an efficient algorithm for cluster architectures that can be used to reduce the overall time of large-scale CCSD(T) calculations to acceptable dimensions.

3.1. Benchmark Calculation for the ^{13}C NMR Chemical Shifts of Benzene. In a previous study it has been demonstrated that methods like CCSD(T) can be used to achieve an accuracy of 2−4 ppm deviation from experiment in the calculation of ^{13}C NMR chemical shifts.[53] While this study included 16 small organic molecules, of which the largest cases were CF_4 and acetone (CH_3COCH_3), the limitations of the serial implementation and the limited computational resources did not allow for the calculation of larger molecules. One example for which accurate benchmark results are of immediate interest is benzene as computational studies, especially applying density functional theory, on all kinds of substituted benzene species are abundant in the literature.[54−56] Thus, the parallel algorithm for the calculation of second-order properties described above has been applied to perform CCSD(T) calculations of the NMR chemical shifts of benzene using various basis sets in order to estimate basis set convergence. Here, the new algorithm allows the use of very large basis sets even for a system with 12 atoms and 42 correlated electrons.

The results including NMR chemical shifts and also zero-point vibrational corrections are given in Table 7. An analysis of the basis-set convergence leads to the conclusion that the Dunning basis sets that have been optimized for energies from post-HF correlation methods and that are fairly diffuse are not very suitable for the calculation of the NMR chemical shifts that probe the electron density closer to the nucleus. Even the Dunning core-valence basis sets that are augmented with tight functions do not perform as well as the corresponding Karlsruhe basis sets[62−65] if one aims at quantitative accuracy in the prediction of NMR chemical shifts.

3.2. The Equilibrium Structure of Ferrocene. Within the last 25 years many attempts were made to determine the structure of ferrocene by applying various quantum-chemical methods.[66−70] A more recent study[71] presented first calculations employing analytical CCSD(T) gradients on this problem using a relatively small basis set and the frozen-core approximation. Up to now, quantum-chemical models have great difficulties to determine the equilibrium metal−ligand distance, a quantity that is not directly accessible to experiment but often used for benchmark studies in the framework of density-functional theory. The structural parameters of ferrocene in its eclipsed (equilibrium) and staggered (saddle point) conformations have been determined using analytic CCSD(T) gradients correlating all 96 electrons with a full triple-ζ quality basis set. Using the cc-pVTZ basis set[33,72] (508 basis functions) one geometry cycle takes about 2.3 days when performing the calculation on 15 nodes. With the cc-pwCVTZ basis set[72,73] (572 basis functions) a geometry cycle takes about 8.8 days using 14 nodes. The results in comparison with previous coupled-cluster studies are presented in Tables 8 and 9. The coupled-cluster results show a relatively pronounced basis set dependence. Quadruple-ζ quality CCSD(T) calculations again correlating all electrons are underway.

4. Conclusions

A detailed analysis of the time-determining steps in CC energy, gradient, and second derivative calculations shows that for almost all practical applications only a few terms completely dominate the overall computation time. This motivates a straightforward strategy for the parallelization of CCSD and CCSD(T) energies, gradients, and second derivatives that has been outlined in this paper. Starting from the highly efficient serial implementation of the ACES II MAB computer code an adaptation for affordable workstation

Table 8. Structure Parameters of the Eclipsed Conformation of Ferrocene[a]

method	no. of basis functions	Fe−C$_5$	Fe−C	C−C	C−H	<C$_5$−H	ref
fc-CCSD(T)/TZ2P+f[b]	373	1.655	2.056	1.433	1.077	1.03	71
CCSD(T)/cc-pVTZ	508	1.639	2.039	1.426	1.075	0.45	this work
CCSD(T)/cc-pwCVTZ	672	1.648	2.047	1.427	1.079	0.52	this work

[a] Bond lengths are given in Å; angles are given in deg. [b]fc (frozen core) denotes that only the 66 valence electrons were correlated.

Table 9. Structure Parameters of the Staggered Conformation of Ferrocene[a]

method	no. of basis functions	Fe−C_5	Fe−C	C−C	C−H	<C_5−H	ref
fc-CCSD(T)/TZ2P+f[b]	373	1.659	2.058	1.432	1.077	1.34	71
CCSD(T)/cc-pVTZ	508	1.642	2.041	1.425	1.075	0.67	this work
CCSD(T)/cc-pwCVTZ	672	1.652	2.050	1.426	1.078	0.61	this work

[a] Bond lengths are given in Å; angles are given in degrees. [b] fc (frozen core) denotes that only the 66 valence electrons were correlated.

clusters has been obtained by parallelizing the most time-consuming steps of the algorithm.

This also includes the calculation of parallel CCSD energies and gradients using unrestricted (UHF) and restricted open-shell (ROHF) Hartree−Fock references, parallel UHF-CCSD(T) energies and gradients, parallel ROHF-CCSD(T) energies as well as parallel equation-of-motion CCSD energies and gradients for closed- and open-shell references.

The central aspect of the implementation presented here is the replication of the cluster amplitudes and the distributed evaluation, storage, and access of the two-electron integrals to arrive at an algorithm for which sufficient local memory and disk space are necessary but which is not dependent on sophisticated high-speed network connections.

Benchmark calculations for systems with up to 1300 basis functions show that the resulting algorithm for energies, gradients, and second derivatives at the CCSD and CCSD(T) level of theory exhibits good scaling with the number of processors as long as the terms that are the time-determining steps in the serial calculation still dominate the overall time in the parallel computation. It is important to note that the communication steps within the algorithm are at no point bottlenecks in the current implementation, even if 16 or more processors are used. Nevertheless, at larger numbers of nodes the algorithm will break down, as steps in the CCSD algorithm that have not been parallelized prevent a better scaling of the overall execution time, especially for small systems and large number of nodes. The current limitation of the parallel implementation becomes obvious for more than 16 processors. However, an analysis of the algorithm leads us to the conclusion that the scaling behavior is much better for larger examples, where the time-determining steps that have been parallelized dominate the overall execution time more strongly.

If a very rough estimate is allowed at this point—implementations of this kind would open the field of application for the CC hierarchy of high accuracy ab initio methods to systems of about 30 atoms in a triple-ζ basis or about 15 atoms in a quadruple-ζ basis. Typical applications would be calculations of the type presented in the last sections of this paper like high accuracy calculations for structures and energies, vibrational frequencies, or properties related to the NMR spectroscopy of molecules with importance for homogeneous catalysis, model systems for biochemistry, or state-of-the-art spectroscopy.

Technical Details. All calculations were carried out on a 16 node single core 3.4 GHz Intel Xeon (EM64T) cluster with 2 MByte L2 Cache and 16 GB DDR-333 RAM on each node. For the network communication a channel bonded Gigabit Ethernet was used. Channel bonding was set up using the two already built-in network interfaces of the compute nodes by using the standard Linux kernel drivers. This resulted in about 50% more network throughput in comparison to one single Gigabit Ethernet connection per node. For the parallel implementation the message passing interface (MPI)[23] is used. The results presented here are obtained by using LAM/MPI.[74,75] All communication in our implementation is done by the MPI_ALLREDUCE subroutine.

Acknowledgment. We thank Dr. Jonas Jusélius (University of Tromsø, Norway) for help with various computational aspects of this work and Professor Kenneth Ruud (University of Tromsø, Norway) for hospitality and helpful discussions. Work in Chemnitz and in Mainz has been supported by the Deutsche Forschungsgemeinschaft (AU 206/1-1 and GA 370/5-1) as well as by the Fonds der Chemischen Industrie.

References

(1) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599.

(2) Bomble, Y. J.; Vázquez, J.; Kállay, M.; Michauk, C.; Szalay, P. G.; Császár, A. G.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2006**, *125*, 064108.

(3) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129.

(4) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.

(5) Heckert, M.; Kállay, M.; Tew, D. P.; Klopper, W.; Gauss, J. *J. Chem. Phys.* **2006**, *125*, 044108.

(6) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

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

(8) Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 214105.

(9) Botschwina, P. *Theor. Chem. Acc.* **2005**, *114*, 350.

(10) Hill, J.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072.

(11) Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Theory Comput.* **2007**, *3*, 1312.

(12) Watts, J. D. *Parallel Computing* **2000**, *26*, 857.

(13) 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 and earlier versions, a package of ab initio programs*; 2006. See http://www.molpro.net (accessed August 2007).

(14) Kállay, M.; Harding, M. E. *Parallel version of the string-based general coupled-cluster program;* 2006. See http://www.mrcc.hu (accessed August 2007).

(15) Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *118*, 7751.

(16) Hirata, S. *J. Phys. Chem. A* **2003**, *107*, 9887.

(17) Aprá, E.; Windus, T. L.; Straatsma, T. P.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Kowalski, K.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.7;* Pacific Northwest National Laboratory: Richland, WA 99352−0999, U.S.A., 2005.

(18) Schmidt, M.; Baldridge, K.; Boatz, J.; Elbert, S.; Gordon, M.; Jensen, J.; Koseki, S.; Matsunaga, N.; Su, K. N. S.; Windus, T.; Dupuis, M.; Montgomery, J. *J. Comput. Chem.* **1993**, *14*, 1347.

(19) *PQS version 3.2;* Parallel Quantum Solutions: 2005. See http://www.pqs-chem.com.

(20) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368.

(21) Olson, R. M.; Schmidt, M. W.; Gordon, M. S.; Rendell, A. P. Enabling the Efficient Use of SMP Clusters: The GAMESS/DDI Approach. In Supercomputing, 2003 ACM/IEEE Conference, Phoenix, AZ, 2003; p 41.

(22) Ford, A. R.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2007**, *28*, 1215.

(23) The MPIForum, MPI: a message passing interface. In Proceedings of the 1993 ACM/IEEE conference on Supercomputing, ACM Press: Portland, OR, U.S.A., 1993.

(24) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; with contributions from: Auer, A. A.; Bernholdt, D.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; Price, D. R.; Ruud, K.; Schiffmann, F.; Tajti, A.; Varner, M. E.; Vázquez, J. including the following integral packages: MOLECULE (J. Almlöf and P. R. Taylor), PROPS (P. R. Taylor), and ABACUS (T. Helgaker, H. J. Aa. Jensen, P. Jørgensen, and J. Olsen). *Aces II Mainz-Austin-Budapest version*; 2007. See http://www.aces2.de (accessed August 2007).

(25) Baumgartner, G.; Auer, A. A.; Bernholdt, D. E.; Bibireata, A.; Choppella, V.; Cociorva, D.; Gao, X.; Harrison, R.; Hirata, S.; Krishanmoorthy, S.; Krishnan, S.; Lam, C.-C.; Nooijen, M.; Pitzer, R.; Ramanujam, J.; Sadayappan, P.; Sibiryakov, A. *Proc. IEEE* **2005**, *93*, 276.

(26) In statements of scaling behavior "*occ*" stands for number of occupied orbitals, and "*vrt*" stands for the number of virtual orbitals. For methods from the coupled cluster hierarchy only the scaling for the time-determining step is given as an estimate for the overall scaling, assuming that the number of virtual orbitals is larger than the number of occupied orbitals.

(27) It should be noted that the size required for storage of the two-electron integrals is of the order of more than 500 GB for a comparable example.

(28) Watts, J. D.; Gauss, J.; Bartlett, R. J. *Chem. Phys. Lett.* **1992**, *200*, 1.

(29) Gauss, J.; Stanton, J. F. *Chem. Phys. Lett.* **1997**, *276*, 70.

(30) Szalay, P. G.; Gauss, J.; Stanton, J. F. *Theor. Chem. Acc.* **1998**, *100*, 5.

(31) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1996**, *104*, 2574.

(32) Gauss, J.; Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *95*, 2623.

(33) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(34) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2975.

(35) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1995**, *103*, 4572.

(36) Rendell, A. P.; Guest, M. F.; Kendall, R. A. *J. Comput. Chem.* **1993**, *14*, 1429.

(37) Auer, A. A.; Baumgartner, G.; Bernholdt, D. E.; Bibireata, A.; Choppella, V.; Cociorva, D.; Gao, X.; Harrison, R.; Krishanmoorthy, S.; Krishnan, S.; Lu, Q.; Lam, C.-C.; Nooijen, M.; Pitzer, R.; Ramanujam, J.; Sadayappan, P.; Sibiryakov, A. *Mol. Phys.* **2006**, *104*, 211.

(38) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718.

(39) Rendell, A. P.; Lee, T. J. *J. Chem. Phys.* **1991**, *94*, 6219.

(40) The quantity $\tilde{t}_{ijk}^{abc}$ arises as a disconnected term in the fifth-order energy correction of the perturbative triples correction.

(41) All calculations for the benzene molecule were carried out at the all electron CCSD(T)/cc-pVQZ geometry (r_{CH}=1.0800 Å, r_{CC}=1.3911 Å), which was taken from ref 76.

(42) Due to machine load and other influences the timings can be assumed to be accurate to a few seconds walltime.

(43) In practice the factorization of the CC equations leads to additional terms that are included in this contraction.

(44) Meyer, W. *J. Chem. Phys.* **1975**, *64*, 1975.

(45) Ahlrichs, R.; Zirz, C. Proceedings of the workshop "Molecular Physics and Quantum Chemistry", Wollongong, 1980.

(46) Ahlrichs, R.; Zirz, C. *Theor. Chim. Acta* **1976**, *36*, 275.

(47) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem. Symp.* **1976**, *10*, 1.

(48) Hampel, C.; Peterson, K.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *192*, 1.

(49) Koch, H.; Christiansen, O.; Kobayashi, R.; Jørgensen, P.; Helgaker, T. *Chem. Phys. Lett.* **1994**, *228*, 233.

(50) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *103*, 3561.

(51) Price, D.; Szalay, P. G.; Harding, M. E.; Vázquez, J.; Stanton, J. F. to be published, 2006.

(52) Fossgård, E.; Ruud, K. *J. Comput. Chem.* **2006**, *27*, 326.

(53) Auer, A. A.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2003**, *118*, 10407.

(54) Dumont, E.; Chaquin, P. *Chem. Phys. Lett.* **2007**, *435*, 354.

(55) Cheng, J.; Zhu, W.; Tang, Y.; Xu, Y.; Li, Z.; Chen, K.; Jiang, H. *Chem. Phys. Lett.* **2006**, *422*, 455.

(56) Heine, T.; Corminboeuf, C.; Grossmann, G.; Haeberlen, U. *Angew. Chem., Int. Ed.* **2006**, *45*, 7292.

(57) Jameson, A. K.; Jameson, C. J. *Chem. Phys. Lett.* **1987**, *134*, 461.

(58) Sundholm, D.; Gauss, J.; Schäfer, A. *J. Chem. Phys.* **1996**, *105*, 11051.

(59) London, F. *J. Phys. Radium* **1937**, *8*, 397.

(60) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(61) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789.

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

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

(64) Gauss, J. *J. Chem. Phys.* **1993**, *99*, 3629.

(65) Auer, A. A. Ph.D. Thesis, Universität Mainz, Mainz, Germany, 2002.

(66) Lüthi, H. P.; Ammenter, J. H.; Almlöf, J.; Fægri, K. *J. Chem. Phys.* **1982**, *77*, 2002.

(67) Klopper, W.; Lüthi, H. P. *Chem. Phys. Lett.* **1996**, *262*, 546.

(68) Koch, H.; Jørgensen, P.; Helgaker, T. *J. Chem. Phys.* **1996**, *104*, 9528.

(69) Lüthi, H. P. *J. Mol. Struct.* (*THEOCHEM*) **1996**, *388*, 299.

(70) Xu, Z.-F.; Xie, Y.; Feng, W.-L.; Schaefer, H. F., III *J. Phys. Chem. A* **2003**, 107, 2716.

(71) Coriani, S.; Haaland, A.; Helgaker, T.; Jørgensen, P. *Chem. Phys. Chem.* **2006**, *7*, 245.

(72) Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **2002**, *117*, 10548.

(73) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.

(74) Burns, G.; Daoud, R.; Vaigl, J. LAM: An Open Cluster Environment for MPI. Proceedings of Supercomputing Symposium, 1994; pp 379−386.

(75) Squyres, J. M.; Lumsdaine, A. A Component Architecture for LAM/MPI. Proceedings, 10th European PVM/MPI Users' Group Meeting, Venice, Italy, 2003; pp 379−387.

(76) Gauss, J.; Stanton, J. F. *J. Phys. Chem. A* **2000**, *104*, 2865.

CT700152C

JCTC Journal of Chemical Theory and Computation

Zn Coordination Chemistry: Development of Benchmark Suites for Geometries, Dipole Moments, and Bond Dissociation Energies and Their Use To Test and Validate Density Functionals and Molecular Orbital Theory

Elizabeth A. Amin*,† and Donald G. Truhlar‡

Department of Medicinal Chemistry, College of Pharmacy, University of Minnesota, 717 Delaware St. SE, Minneapolis, Minnesota 55414-2959, and Department of Chemistry, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota 55455-0431

Received August 17, 2007

Abstract: We present nonrelativistic and relativistic benchmark databases (obtained by coupled cluster calculations) of 10 Zn−ligand bond distances, 8 dipole moments, and 12 bond dissociation energies in Zn coordination compounds with O, S, NH_3, H_2O, OH, SCH_3, and H ligands. These are used to test the predictions of 39 density functionals, Hartree−Fock theory, and seven more approximate molecular orbital theories. In the nonrelativisitic case, the M05-2X, B97-2, and mPW1PW functionals emerge as the most accurate ones for this test data, with unitless balanced mean unsigned errors (BMUEs) of 0.33, 0.38, and 0.43, respectively. The best local functionals (i.e., functionals with no Hartree−Fock exchange) are M06-L and τ-HCTH with BMUEs of 0.54 and 0.60, respectively. The popular B3LYP functional has a BMUE of 0.51, only slightly better than the value of 0.54 for the best local functional, which is less expensive. Hartree−Fock theory itself has a BMUE of 1.22. The M05-2X functional has a mean unsigned error of 0.008 Å for bond lengths, 0.19 D for dipole moments, and 4.30 kcal/mol for bond energies. The X3LYP functional has a smaller mean unsigned error (0.007 Å) for bond lengths but has mean unsigned errors of 0.43 D for dipole moments and 5.6 kcal/mol for bond energies. The M06-2X functional has a smaller mean unsigned error (3.3 kcal/mol) for bond energies but has mean unsigned errors of 0.017 Å for bond lengths and 0.37 D for dipole moments. The best of the semiempirical molecular orbital theories are PM3 and PM6, with BMUEs of 1.96 and 2.02, respectively. The ten most accurate functionals from the nonrelativistic benchmark analysis are then tested in relativistic calculations against new benchmarks obtained with coupled-cluster calculations and a relativistic effective core potential, resulting in M05-2X (BMUE = 0.895), PW6B95 (BMUE = 0.90), and B97-2 (BMUE = 0.93) as the top three functionals. We find significant relativistic effects (~0.01 Å in bond lengths, ~0.2 D in dipole moments, and ~4 kcal/mol in Zn−ligand bond energies) that cannot be neglected for accurate modeling, but the same density functionals that do well in all-electron nonrelativistic calculations do well with relativistic effective core potentials. Although most tests are carried out with augmented polarized triple-ζ basis sets, we also carried out some tests with an augmented polarized double-ζ basis set, and we found, on average, that with the smaller basis set DFT has no loss in accuracy for dipole moments and only ~10% less accurate bond lengths.

1. Introduction

Zinc is an essential element for humans, primarily because it serves as a cofactor for a very large number of enzyme reactions[1,2] (it is the second most abundant transition metal cation in biology[3]), and it is technologically important in ZnO photoluminescent materials and nanoparticles (quantum dots).[4] Zinc-binding proteins that perform essential functions in a variety of species, and for which accurate active-site modeling parameters are needed, include insulin, metallothionein, DNA topoisomerase, phosphotriesterase (an

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

† Department of Medicinal Chemistry.

‡ Department of Chemistry.

10.1021/ct700205n CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/30/2007

enzyme that hydrolyzes organophosphorus compounds like sarin[5]), zinc-finger proteins, matrix metalloproteinases (MMPs), the anthrax toxin lethal factor, alcohol dehydrogenase, human carbonic anhydrase,[6] cytidine deaminase, β-lactamases, and copper-zinc dismutase. In metalloproteins, Zn often functions as a Lewis acid, with catalysis occurring in its first coordination shell, or by electrostatically stabilizing reactants or intermediates in the active site. Zinc fluorescent sensors used to monitor labile Zn^{2+} [7] and chelating agents used in froth flotation[8] also operate by inner-shell coordination. Geometries, dipole moments, and bond dissociation energies of Zn coordination compounds are thus critical parameters for reliable simulations of biological function[2,9−30] and technological applications of Zn chemistry. Density functional theory (DFT) is a very promising electronic structure calculation tool for obtaining such parameters,[31−33] but systematic validation studies, which have not yet been reported, are required to understand the reliability of DFT calculations for Zn binding as well as the reliability of simpler semiempirical calculations that are often used because they are faster and less expensive.

The present paper presents nonrelativistic and relativistic validation suites of bond distances for 10 zinc−ligand complexes, dipole moments for 8 Zn model compounds, and bond dissociation energies (BDEs) for 12 zinc compounds. In obtaining all three parameter sets, we chose model compounds with Zn bonds to N, O, and S, which are the three most common first-coordination-shell atoms in zinc enzymes.[25,34−37] Although our first priority was to test model compounds representing biozinc centers, we also include one compound with Zn bonds to H, which occur at defect sites in ZnO crystals[38] and in hydrogen-doped zinc oxide thin films.[39] These data are then used to test practical electronic structure methods of both fundamental types: density functional theory (DFT)[40] and wave function theory (WFT).[41,42]

DFT methods evaluated with a nonrelativistic Hamiltonian include the five most accurate functionals[43−54] for metal−ligand bond energies in a recent study[55] of 21 metal−ligand complexes with 57 different density functionals and the two most accurate functionals[48,51,55,56] overall in that study (based on the 21 metal−ligand bond energies, 8 transition-metal dimer bond energies, 6 representative main-group atomization energies, 7 ionization potentials, 13 metal−ligand bond lengths, and 8 transition metal dimer bond lengths) plus five density functionals developed subsequently[57−60] and 26 other popular and representative density functionals[43−46,48,49,51,54,56,61−78] of various types. The NDO methods evaluated for comparison with DFT calculations are AM1,[9,79−81] MNDO,[82,83] MNDO/d,[84,85] PM3,[86] PM3(tm),[86−90] and the newer PM6 method.[91] (The PM3(tm) method is tested only for geometries and bond energies because it was not parametrized for dipole moments[88] and is specifically stated by its developers not to be used for that purpose.) The IEHT method that we test is self-consistent-charge density-functional tight-binding (SCC-DFTB).[92−94] For the nonrelativistic evaluations, we test the following: 39 density functionals, both local and nonlocal; ab initio Hartree−Fock (HF) theory; six semiempirical molecular orbital methods of the neglect of differential overlap (NDO) variety; and one iterative extended Hückel theory (IEHT) method, which is also called a tight-binding method. These calculations involve testing nonrelativistic DFT calculations against nonrelativistic benchmarks. We then incorporate a relativistic effective core potential (ECP)[95] for Zn into the top ten methods resulting from this analysis and test these relativistic DFT calculations against new benchmarks that also incorporate relativistic effects via this ECP.

2. Data Sets and Computational Methods

2.1. Basis Sets. We use two basis sets in this work, and we will denote them B1 and B2. Basis set B1 is used for nonrelativistic DFT calculations. In this basis, Zn is represented by the 6-311+G(d,p) basis set of in *Gaussian 03*,[96] which is constructed from the earlier work of Wachters[97] and Hay[98] as modified by Raghavachari and Trucks,[99] who described both an "*spd*" basis and an "*spdf*" basis. The basis used here is the *spd* basis further polarized with a single *f* set with an exponent[99] of 1.62. The final Zn basis for B1 consists of a 15*s*11*p*6*d*1*f* primitive basis contracted to 10*s*7*p*4*d*1*f*, with the outer functions uncontracted. For H, C, N, O, and S in Series A, the basis set is MG3S′, which denotes MG3S[100] with oxygen *f* functions removed to decrease computation time. Thus the MG3S′ basis set is 6-311+G(2df,2p)[100−102] for H, C, and N, 6-311+G(2d)[102−104] for O, and the same as in G3Large[105] without core polarization functions for S.

Basis set B2 is used for relativistic DFT calculations, and for both nonrelativistic and relativistic WFT calculations (in particular, for CCSD(T) and CCSD benchmarks and for MP2 calculations used for extrapolation). We augment the Zn basis by adding two additional *f* functions specified by Raghavachari and Trucks,[99] with exponents of 0.486 and 5.40, and we restore the MG3S *f* functions to oxygen.

In both basis sets we use five spherical harmonic basis functions for *d* sets and seven spherical harmonic basis functions for *f* sets.

A smaller basis set will be discussed briefly in section 4.1.

2.2. Core Orbitals and Relativistic Effects. Relativistic effects may be divided into scalar relativistic effects and vector effects.[106] The most important vector relativistic effect is spin−orbit coupling, but, except for ZnH, all Zn-containing species treated in the present article are closed-shell singlets for which spin−orbit coupling vanishes. Spin−orbit coupling also vanishes for ZnH because it is a $^2\Sigma^+$ state. Spin−orbit coupling is nonzero for O and S and was accounted for in all relativistic calculations (benchmarks and more approximate methods) by subtracting 0.02 kcal/mol (O) and 0.56 kcal/mol (S) from the calculated bond dissociation energies for processes that respectively produce O and S, based on atomic energy levels.[107]

Scalar relativistic effects are very important for 5*d* transition metals, important for 4*d* transition metals, and "small" but not negligible for 3*d* transition metals like Zn. For 3*d* and 4*d* transition metals, an adequate way to include scalar relativistic effects in either WFT or DFT is to replace the inner core orbitals by a relativistic effective core potential. (A recent study validating these procedures for PdCO may be consulted for further discussion.[108]) In the present article, the inner core is defined as the next-to-largest noble gas core, which is sometimes called the "small core" prescription (thus the inner core for Zn has 10 electrons, whereas a large-core prescription (not used here) would treat 18 electrons as the core). In all WFT and DFT relativistic calculations, scalar relativistic effects at Zn are included by replacing the inner core electrons by the quasirelativistic multielectron-fit (MEFIT, *R*) pseudopotential of Preuss et al.[95] Note that we do *not* use the basis set developed[95] for use with this pseudopotential; all relativistic calculations in the present article use the B2 basis set.

In the present nonrelativistic calculations, all electrons are included explicitly with no effective core potential. In the nonrelativistic DFT calculations, all electrons are correlated. The nonrelativistic WFT calculations consist of an uncorrelated Hartree−Fock step followed by a correlated post-Hartree−Fock step. In the latter, the inner core (10 electrons) is not correlated. In extensive tests not presented here, we found that using the large-core prescription for the post-Hartree−Fock step gives significantly different geometries and should not be trusted.

2.3. Benchmark Databases. With the exception of ZnH_2, for which the gas-phase equilibrium internuclear distance r_e has been obtained by high-resolution infrared emission spectroscopy,[109,110] experimental bond lengths and dipole moments have not yet been published for the model compounds in this study. Therefore our first goal is to assemble benchmark data sets. For this purpose, we first consider a set of ten Zn model compounds. For eight of these, in particular, ZnO, ZnS, $Zn(NH_3)^{2+}$ $Zn(SCH_3)^+$, $Zn(H_2O)^{2+}$, $Zn(OH)_2$, $ZnOH^+$, and ZnH_2, coupled cluster theory[111] with single and double excitations[112] and quasiperturbative triple excitations (CCSD(T))[113] with basis B2 is applied to obtain the best estimates of accurate geometric parameters. The CCSD(T) method has been shown in multiple studies[114−117] to reliably and accurately reproduce experimental geometries for small transition-metal model compounds similar to those examined here when the basis set is large enough. For nonrelativistic benchmarks, the next-to-largest noble gas cores were frozen in these post-HF calculations; the relativistic best estimates use the quasirelativistic multielectron-fit (MEFIT,*R*) pseudopotential of Preuss et al. on Zn.[95] The Zn−ligand equilibrium bond-distance values obtained in these ways are shown in Table 1.

The two remaining model compounds, $Zn(NH_3)_3{}^{2+}$ and $Zn(NH_3)_4{}^{2+}$, in the geometry database were too large for CCSD(T)/B2 optimizations; therefore, our best estimates in these cases were obtained by extrapolation from MP2/B2 calculations. In particular, we noted that increasing the level of calculation from MP2/B2 to CCSD(T)/B2 for $Zn(NH_3)^{2+}$ uniformly increases the bond length, as shown in Table 2. It increases the Zn−N bond distance by 0.020 Å (nonrelativistic) and 0.022 Å (relativistic); we therefore increased the MP2/B2 calculated Zn−N r_e values by these amounts for $Zn(NH_3)_3{}^{2+}$ and $Zn(NH_3)_4{}^{2+}$ to obtain the best estimates in Table 1.

Table 1. Best Estimates of Nonrelativistic and Relativistic Zn−Ligand Bond Distances Obtained by CCSD(T)/B2

compound	distance	r_e (Å), nonrel	r_e (Å), rel
ZnO	Zn−O	1.721	1.710
ZnS	Zn−S	2.077	2.067
$Zn(NH_3)^{2+}$	Zn−N	1.955	1.939
$Zn(H_2O)^{2+}$	Zn−O	1.868	1.852
$ZnOH^+$	Zn−O	1.764	1.757
$Zn(SCH_3)^+$	Zn−S	2.181	2.170
$Zn(OH)_2$	Zn−O	1.779	1.767
ZnH_2	Zn−H	1.544	1.528[a]
$Zn(NH_3)_3{}^{2+}$	Zn−N	2.016[b]	2.005[b]
$Zn(NH_3)_4{}^{2+}$	Zn−N	2.072[b]	2.063[b]

[a] For comparison, the experimental r_e value is 1.524 for ZnH_2.[110] [b] Estimated by extrapolation; cf. section 2.3.

Table 2. Comparison of Relativistic and Nonrelativistic MP2/B2 and CCSD(T)/B2 Bond Distances

compound	distance	Δr_e (Å), nonrel[a]	Δr_e (Å), rel[a]
ZnO	Zn−O	0.042	0.046
ZnS	Zn−S	0.028	0.030
$Zn(NH_3)^{2+}$	Zn−N	0.022	0.020
$Zn(H_2O)^{2+}$	Zn−O	0.010	0.010
$ZnOH^+$	Zn−O	0.015	0.017
$Zn(SCH_3)^+$	Zn−S	0.040	0.040
$Zn(OH)_2$	Zn−O	0.020	0.021
ZnH_2	Zn−H	0.025	0.025

[a] $\Delta r_e = [r_e(\text{CCSD(T)/B2}) - r_e(\text{MP2/B2})]$.

Table 3. Nonrelativistic and Relativistic Best Estimates of Dipole Moments Calculated by the CCSD/B2 Method

	dipole moment (D)[a]	
compound	nonrelativistic	relativistic
ZnO	5.69	5.50
ZnS	5.73	5.47
$Zn(NH_3)^{2+}$	1.07	1.27
$Zn(SCH_3)^+$	3.95	3.59
$Zn(H_2O)^{2+}$	0.27	0.39
$ZnOH^+$	4.51	4.27
$Zn(NH_3)_2(OH)_2$	3.86	3.80
$Zn(NH_3)(OH)^+$	7.40	7.34

[a] 1 D ≡ 1 Debye. For ions, the origin is at the center of mass of the nuclei.

Best estimates of dipole moments were obtained by CCSD/B2 optimizations on an eight-compound data set: ZnO, ZnS, $Zn(NH_3)^{2+}$ $Zn(SCH_3)^+$, $Zn(H_2O)^{2+}$, $ZnOH^+$, $Zn(NH_3)_2(OH)_2$, and $Zn(NH_3)(OH)^+$ (Table 3). Benchmarks for Zn−ligand bond dissociation energies were obtained for a 12-compound data set: ZnO, ZnS, $Zn(NH_3)^{2+}$, $Zn(OH)_2$, ZnH_2, $Zn(SCH_3)^+$,$Zn(H_2O)^{2+}$, $ZnOH^+$, $Zn(NH_3)_2(OH)_2$, $Zn(NH_3)(OH)^+$, $Zn(NH_3)_3{}^{2+}$, and $Zn(NH_3)_4{}^{2+}$ (Table 4). For molecules that include ammonia groups in addition to other ligands, the Zn−N BDE is the one considered, as shown in the last two rows of Table 4. Bond dissociation energies for four of these molecules, $Zn(NH_3)(OH)^+$, $Zn(NH_3)_2(OH)_2$, $Zn(NH_3)_3{}^{2+}$, and $Zn(NH_3)_4{}^{2+}$, were again extrapolated from MP2/B2 calculations as they proved too large for CCSD(T) optimizations. Energies for $Zn(NH_3)_3{}^{2+}$ and $Zn(NH_3)_4{}^{2+}$ were estimated based on the 1.29 kcal/mol (nonrelativistic) and 1.42 kcal/mol (relativistic) decrease in BDE for $Zn(NH_3)^{2+}$ from the MP2 to the CCSD(T) level (Table 5). The mean decrease in

Table 4. Best Estimates of Nonrelativistic and Relativistic Zn−Ligand Bond Dissociation Energies (BDEs in kcal/mol) Obtained by CCSD(T)/B2

compound	dissociation products	BDE (kcal/mol), nonrel	BDE (kcal/mol), rel
ZnO	Zn, O	83.51	80.30
ZnS	Zn, S	61.92	58.00
$Zn(NH_3)^{2+}$	Zn^{2+}, NH_3	129.00	134.15
$Zn(H_2O)^{2+}$	Zn^{2+}, H_2O	96.83	99.78
$ZnOH^+$	Zn^{2+}, OH^-	428.18	435.66
$Zn(SCH_3)^+$	Zn^{2+}, SCH_3^-	420.60	433.84
$Zn(OH)_2$	$ZnOH^+$, OH^-	256.81	258.71
ZnH_2	ZnH, H	78.72	78.90
$Zn(NH_3)_3^{2+}$	$Zn(NH_3)_2^{2+}$, NH_3	61.74[a]	60.67[a]
$Zn(NH_3)_4^{2+}$	$Zn(NH_3)_3^{2+}$, NH_3	46.81[a]	46.04[a]
$Zn(NH_3)_2(OH)_2$	$Zn(NH_3)(OH)_2$, NH_3	10.31[a]	8.87[a]
$Zn(NH_3)(OH)^+$	$ZnOH^+$, NH_3	79.31[a]	81.14[a]

[a] Estimated by extrapolation; cf. section 2.3.

Table 5. Comparison of Relativistic and Nonrelativistic MP2/B2 and CCSD(T)/B2 Bond Dissociation Energies (BDEs)

compound	dissociation products	Δ_{BDE} (kcal/mol), nonrel[a]	Δ_{BDE} (kcal/mol), rel[a]
ZnO	Zn, O	−31.75	−31.30
ZnS	Zn, S	−16.82	−17.17
$Zn(NH_3)^{2+}$	Zn^{2+}, NH_3	−1.29	−1.42
$Zn(H_2O)^{2+}$	Zn^{2+}, H_2O	−0.94	−1.02
$ZnOH^+$	Zn^{2+}, OH^-	0.34	0.19
$Zn(SCH_3)^+$	Zn^{2+}, SCH_3^-	−1.63	1.73
$Zn(OH)_2$	$ZnOH^+$, OH^-	−2.41	−2.79
ZnH_2	ZnH, H	1.02	0.77

[a] Δ_{BDE} = [BDE (CCSD(T)/B2) − BDE (MP2/B2)].

Table 6. Summary of the DFT Methods Evaluated in This Study[a]

functional	type	X	τ in E or C?	refs
B1LYP	HGGA	25	neither	48,64,71
B3LYP	HGGA	20	neither	48,67,68
B97-2	HGGA	21	neither	54
BLYP	GGA	0	neither	48,64
BP86	GGA	0	neither	63,64
BVWN5	GGE	0	neither	62,64
G96HLYP	GGSC	0	neither	48,55,56
G96LYP	GGA	0	neither	48,56
G96LYP1M[b]	GGSC	0	neither	118
G96VWN5	GGE	0	neither	56,62
τ-HCTH	MGGA	0	exchange	73,75
M05	HMGGA	28	both	57
M05-2X	HMGGA	56	both	58
M06	HMGGA	27	both	60
M06-2X	HMGGA	54	both	60
M06-L	MGGA	0	both	59
MOHLYP	GGSC	0	neither	48,49,55
MPW1B95	HMGGA	31	correlation	51,69,77
MPW1KCIS	HMGGA	15	correlation	43−45,51,52
mPW1PW[c]	HGGA	25	neither	51,65
mPWLYP	GGA	0	neither	48,51
MPWLYP1M	HGGA	5	neither	48,51,55
mPWVWN5	GGE	0	neither	51,62
O3LYP	HGGA	11.61	neither	48−50
OLYP	GGA	0	neither	48,49
OPWL	GGE	0	neither	49,66
OV5LYP	HGGA	0	neither	48,49,62
OVWN5	GGE	0	neither	49,62
PBEh[d]	HGGA	25	neither	70,74
PBELYP	HGGA	0	neither	48,70
PBEVWN5	GGE	0	neither	62,70
PW6B95	HMGGA	28	correlation	78
SVWN5	LSDA	0	neither	61,62
TPSS	MGGA	0	both	46
TPSS1KCIS	HMGGA	13	both	43−47
TPSSh	HMGGA	10	both	53
TPSSVWN5	GGE	0	exchange	46,62
VSXC[e]	MGGA	0	both	72
X3LYP	HGGA	21.8	neither	38,44,45,68

[a] GGA: generalized-gradient approximation; GGE: generalized-gradient exchange; GGSC: generalized-gradient exchange with scaled correlation; HGGA: hybrid GGA; HMGGA: hybrid meta GGA; LSDA: local spin density approximation; MGGA: meta GGA; X = percentage of Hartree−Fock exchange; E = exchange; C = correlation. [b] G96LYP1M is like G98HLYP except that the gradient correction to the correlation energy is multiplied by 0.54 instead of 0.50. [c] Same as mPWO, mPW1PW91, and MPW25. [d] Same as PBE0 and PBE1PBE. [e] Same as VS98.

BDEs for $Zn(NH_3)^{2+}$, $Zn(OH)_2$, $Zn(H_2O)^{2+}$, and $ZnOH^+$ from MP2/B2 to CCSD(T)/B2 was 1.24 kcal/mol (nonrelativistic) and 1.36 kcal/mol (relativistic); we therefore estimate CCSD(T) BDEs for $Zn(NH_3)(OH)^+$ and $Zn(NH_3)_2(OH)_2$ based on these values. BDEs for the remaining eight molecules were obtained from CCSD(T)/B2 optimizations.

2.4. Density Functionals. The properties of the density functionals we tested are given in Table 6. The five most accurate functionals in ref 55 for metal−ligand bond energies were TPSS1KCIS,[43−47] O3LYP,[48−50] MPW1KCIS,[43−45,51,52] TPSSh,[53] and B97-2.[54] The two most accurate functionals (overall) in ref 55 were G96LYP[48,56] and MPWLYP1M;[48,51,55] BLYP[48,64] and MOHLYP[48,49,55] were also among the five best functionals overall and are also tested here. The new functionals we test are as follows: M05,[57] M05-2X,[58] M06-L[59], M06,[60] M06-2X,[61] and G96LYP1M.[118] In addition to these, we tested an assortment of popular functionals with varying performance ranges.[43−46,48,49,51,55,56,61−78]

The functional set tested here comprises 1 local spin density approximation (LSDA), 5 generalized-gradient approximation (GGA), 7 generalized-gradient exchange (GGE), 3 GGE with scaled correlation (GGSC), 10 hybrid GGA (HGGA), 9 hybrid meta GGA (HMGGA), and 4 meta GGA (MGGA) methods (see Table 6). LSDA functionals depend on the spin densities; GGA functionals depend on the gradient of the spin densities as well as the spin densities themselves; HGGA functionals depend on the percentage of Hartree−Fock (HF) exchange, the density gradients, and the spin densities; MGGA functionals depend on the spin kinetic energy densities τ_σ, the spin density gradients, and the spin densities; and HMGGA functionals depend on τ_σ, HF exchange, the density gradients, and the spin densities. GGE methods combine GGA exchange with LSDA correlation, and in GGSC, a relatively new approach,[55] the Kohn−Sham operator is defined by

$$F = F^{SE} + F^{GCE} + F^{LC} + (Y/100)F^{GCC} \tag{1}$$

where F^{SE} is the Slater local exchange functional,[55] F^{GCE} is the gradient correction to the LSDA exchange, F^{LC} is the LSDA correlation functional, and F^{GCC} is the gradient correction to the LSDA correlation, and Y is the percentage of the gradient correction to correlation that is included. Here we set $Y = 50$ in two of the GGSC functionals in this study, G96HLYP and MOHLYP, as was done in previous work,[55] and we set $Y = 54$ in another.[118] For each theory level we specify the percentage X of Hartree−Fock exchange, and whether τ_σ is included in the exchange and/or correlation functionals.

2.5. Computational Details. CCSD, CCSD(T), DFT, AM1, PM3, and MNDO calculations were carried out using *Gaussian 03* or a locally modified version of *Gaussian 03*[95] on the Minnesota Supercomputing Institute core resources and on an Alienware MJ-12 dual-CPU workstation running under the SUSe Linux Professional 9.3 operating system. MNDO/d and PM3(tm) calculations were obtained on the Alienware MJ-12 with the SPARTAN '02 and '04 Linux software packages.[119] PM6 calculations were performed using MOPAC 2007[120] on Alienware Area-51m and Alienware Sentia machines running Windows XP. SCC-DFTB calculations were done using DFTB/DYLAX[92−94] on the Minnesota Supercomputing Institute core resources.

3. Results

All nonrelativistic DFT methods and HF theory were tested using the B1 basis set against nonrelativistic B2 benchmark values; the top ten functionals resulting from this analysis were then evaluated using the aforementioned pseudopotential[95] against relativistic B2 benchmark values.

3.1. Nonrelativistic Calculations. We first compare the nonrelativistic geometric parameters, dipole moments, and bond dissociation energies obtained by the 39 chosen DFT functionals, HF theory, the six NDO methods, and SCC-DFTB to the nonrelativistic benchmark values we reported in section 2. We included DFT levels with a fairly wide variation in Hartree−Fock exchange, from 0 to 56% as well as HF theory with 100% Hartree−Fock exchange. The quality of our results was evaluated by mean unsigned errors (MUEs) representing the average absolute deviations from calculated benchmark values and also by mean signed errors (MSEs) used to detect systematic error. The mean unsigned errors for DFT and Hartree−Fock Zn−ligand equilibrium bond distances are reported in Table 7, and Table 8 gives analogous results for semiempirical molecular orbital theory. Tables 9 and 10 give the mean unsigned errors for model-compound dipole moments, and Tables 11 and 12 list MUEs for bond dissociation energies. Corresponding MSEs are provided in the Supporting Information.

The balanced mean unsigned error (BMUE) is a unitless quantity that normalizes MUEs for each parameter against the average error over all methods for that parameter and thus serves as a valuable criterion to evaluate the overall performance of each technique

$$\begin{aligned}\mathrm{BMUE} = \{[&\mathrm{MUE(in\ \AA)/AMUE(in\ \AA)}] + \\ &[\mathrm{MUE(in\ D)/AMUE(in\ D)}] + \\ &[\mathrm{MUE(in\ kcal/mol)/AMUE(in\ kcal/mol)}]\}/3 \quad (2)\end{aligned}$$

where AMUE is the average mean unsigned error, i.e., the mean of all MUEs for bond distances (in Å), dipole moments (in D), or bond dissociation energies (in kcal/mol). Table 13 gives nonrelativistic BMUEs for all methods except PM3(tm), which is unsuitable[88] for transition-metal dipole moment calculations and which returned very large errors in dipole moments for our compound set; and SCC-DFTB, which gave enormously inaccurate bond dissociation energies. For these reasons PM3 (tm) and SCC-DFTB were not included in calculating AMUEs.

Table 7. Mean Unsigned Errors (MUEs) in DFT and HF Zn−Ligand Bond Distance for Ten Zinc−Ligand Complexes (Nonrelativistic)[a]

functional	MUE (Å)	functional	MUE (Å)
X3LYP	0.0069	M06-2X	0.0169
PW6B95	0.0072	MPWLYP1M	0.0175
M05-2X	0.0078	G96LYP	0.0177
B3LYP	0.0080	mPWLYP	0.0205
MPW1KCIS	0.0080	OLYP	0.0215
B1LYP	0.0084	BLYP	0.0223
mPW1PW	0.0089	HF	0.0224
PBEh	0.0089	G96LYP1M	0.0236
PB86	0.0090	G96HLYP	0.0241
B97-2	0.0090	OV5LYP	0.0244
TPSSh	0.0094	PBELYP	0.0291
TPSS1KCIS	0.0097	TPSSVWN5	0.0309
MPW1B95	0.0105	G96VWN5	0.0318
M06-L	0.0109	OVWN5	0.0355
TPSS	0.0113	OPWL	0.0358
τ-HCTH	0.0133	mPWVWN5	0.0359
O3LYP	0.0139	PBEVWN5	0.0368
M05	0.0147	BVWNS	0.0377
M06	0.0147	SVWN5	0.0410
VSXC	0.0151	MOHLYP	0.0769

[a] Nonrelativistic DFT/B1 tested against nonrelativistic CCSD(T)/B2.

Table 8. Mean Unsigned Errors (MUEs) in NDO and SCC-DFTB Zn−Ligand Bond Distance for Ten Zinc−Ligand Complexes (Nonrelativistic)

method	MUE (Å)
SCC-DFTB	0.043
PM3(tm)	0.061
AM1	0.063
PM3	0.069
PM6	0.077
MNDO(d)	0.078
MNDO	0.082

Table 14 shows results for a smaller basis set discussed in section 4.1

3.2. Relativistic Calculations. We used the same error measures for the relativistic comparisons. Table 15 lists relativistic BMUEs for the top ten functionals from Table 13, and tables analogous to Tables 7, 9, and 11, but for relativistic calculations, are given in the Supporting Information.

4. Discussion

4.1. Nonrelativistic Tests. The X3LYP functional shows the best performance for bond lengths in the nonrelativistic calculations, followed by PW6B95, M05-2X, and B3LYP. These methods all have $20 \leq X \leq 56$. The next two functionals in the ranking have $X = 15$ and $X = 25$. The

Table 9. Mean Unsigned Errors (MUEs) in DFT and HF Dipole Moment for Eight Zinc−Ligand Complexes (Nonrelativistic)[a]

functional	MUE (D)	functional	MUE (D)
M05-2X	0.19	OV5LYP	0.47
B97-2	0.26	OLYP	0.51
M05	0.27	TPSS	0.54
mPW1PW	0.28	TPSSVWN5	0.57
MPW1B95	0.29	MPW1KCIS	0.61
PBEh	0.29	VSXC	0.63
PW6B95	0.33	G96VWN5	0.63
O3LYP	0.36	G96HLYP	0.70
M06-2X	0.37	G96LYP1M	0.70
TPSSh	0.38	mPWVWN5	0.72
TPSS1KCIS	0.39	BP86	0.73
B1LYP	0.39	MPWLYP1M	0.75
OVWN5	0.39	PBEVWN5	0.75
OPWL	0.39	PBELYP	0.81
MOHLYP	0.40	BVWN5	0.82
τ-HCTH	0.41	BLYP	0.85
X3LYP	0.43	mPWLYP	0.85
M06-L	0.43	HF	0.87
B3LYP	0.44	SVWN5	0.89
M06	0.45	G96LYP	0.90

[a] Nonrelativistic DFT/B1 tested against nonrelativistic CCSD/B2.

Table 10. Mean Unsigned Errors (MUEs) in NDO and SCC-DFTB Dipole Moment for Eight Zinc−Ligand Complexes

method	MUE (D)
AM1	0.79
PM3	0.86
PM6	1.15
MNDO	1.28
MNDO(d)	1.29
SCC-DFTB	1.45

Table 11. Mean Unsigned Errors (MUEs) in DFT and HF Bond Dissociation Energy (BDE) for Twelve Zinc−Ligand Complexes (Nonrelativistic)[a]

functional	MUE (kcal/mol)	functional	MUE (kcal/mol)
M06-2X	3.30	G96HLYP	7.27
M05	3.34	TPSSh	7.28
M06	3.80	G96LYP1M	7.33
B97-2	3.94	OV5LYP	7.47
B1LYP	4.26	mPWVWN5	7.50
M05-2X	4.30	OLYP	7.74
MPW1B95	5.16	OVWN5	7.87
mPW1PW	5.24	OPWL	7.88
PW6B95	5.26	PBEVWN5	7.90
M06-L	5.40	VSXC	8.04
B3LYP	5.41	TPSS	8.41
O3LYP	5.51	MPWLYP1M	8.49
X3LYP	5.64	BLYP	8.67
TPSSVWN5	5.97	PBELYP	8.89
PBEh	5.99	BP86	9.06
TPSS1KCIS	6.51	mPWLYP	9.49
MPW1KCIS	6.59	G96LYP	9.50
G96VWNS	6.66	MOHLYP	13.04
τ-HCTH	6.73	HF	15.02
BVWN5	6.99	SVWN5	22.06

[a] Nonrelativistic DFT/B1 tested against nonrelativistic CCSD(T)/B2.

Table 12. Mean Unsigned Errors (MUEs) in NDO and SCC-DFTB Bond Dissociation Energy (BDE) for Twelve Zinc−Ligand Complexes

method	MUE (kcal/mol)
PM6	13.72
PM3	20.43
AM1	27.37
MNDO(d)	28.19
MNDO	31.26
PM3(tm)	84.96
SCC-DFTB	302.13

first 25 functionals have MUE < 0.022 Å, whereas Table 8 shows that even the best of the semiempirical molecular orbital methods has MUE = 0.043 Å.

It is interesting to compare the results in Table 7 to a previous study[121] of bond lengths in van der Waals complexes that included results for Zn_2, ZnNe, ZnAr, and ZnKr by 19 different density functionals. If one computes the mean unsigned errors on those four compounds, the best result (0.28 Å) was obtained by using M05-2X. (The mean unsigned error is larger than the typical value in the present work because the van der Waals complexes have flatter potentials than those for the covalent and coordinate covalent bonds studied here.) Seven other functionals included in that study are also included here, and their mean unsigned errors for the four Zn-containing van der Waals molecules are (in Å) as follows: PW6B95, 0.38; PBEh, 0.41; MPW1B95, 0.44; M05, 0.45; TPSSh, 0.59; TPSS, 0.59; and mPW1PW, 0.67. It is encouraging that the M05-2X and PW6B95 density functionals perform relatively well in both the previous and the present studies.

Table 9 shows that the M05-2X functional predicts the most accurate nonrelativistic dipole moments by a large margin. The B97-2, M05, mPW1PW, MPW1B95, and PBEh functionals are in second through sixth place, with MUE ≤ 0.29 D. In contrast the top semiempirical molecular orbital method, AM1, has MUE = 0.79 D (Table 10). We note that SCC-DFTB is more expensive than NDO methods, but—despite its name—its performance is more similar to other NDO methods than to DFT.

The M06-2X functional is the top DFT method for nonrelativistic bond dissociation energies (Table 11), followed closely by the M05, M06, B97-2, B1LYP, and M05-2X functionals. Once again the NDO methods prove inferior, with PM6 as the best method in this class with MUE = 13.72 kcal/mol. Most notably, SCC-DFTB rendered highly inaccurate BDEs for our compound set, with MUE = 302.13 kcal/mol. In the remainder of the discussion we focus on the unitless balanced MUE (BMUE), which takes account of all three parameters examined here: bond distances, dipole moments, and bond dissociation energies.

Overall, for nonrelativistic BMUE, Table 13 shows that DFT methods perform significantly better than NDO methods and SCC-DFTB for the model Zn model compounds in this study. BMUEs for DFT methods ranged from 0.333 to 1.684,

Table 13. Balanced Mean Unsigned Errors (BMUEs, Unitless) for Three Databases of Zn−Ligand Compounds for DFT, HF, and NDO Methods (Nonrelativistic)[a]

functional	BMUE	functional	BMUE
M05-2X	0.333	OVWN5	0.900
B97-2	0.376	OPWL	0.904
mPW1PW	0.426	G96LYP1M	0.909
M05	0.437	G96HLYP	0.912
PW6B95	0.438	G96VWN5	0.953
MPW1B95	0.450	BLYP	1.00
B1LYP	0.451	G96LYP	1.01
PBEh	0.456	mPWLYP	1.02
X3LYP	0.496	mPWVWN5	1.08
B3LYP	0.510	PBELYP	1.09
M06-2X	0.518	PBEVWN5	1.12
TPSS1KCIS	0.535	BVWN5	1.14
M06-L	0.541	HF	1.22
O3LYP	0.542	MOHLYP	1.58
TPSSh	0.550	SVWN5	1.68
M06	0.551	PM3	1.96
τ-HCTH	0.601	PM6	2.02
MPW1KCIS	0.644	AM1	2.06
TPSS	0.701	MNDO(d)	2.56
VSXC	0.784	MNDO	2.69
OLYP	0.786		
OV5LYP	0.796		
BP86	0.800		
TPSSVWN5	0.885		
MPWLYP1M	0.895		

[a] Nonrelativistic DFT/B1 and other methods tested against nonrelativistic CCSD(T)/B2 for geometries and bond dissociation energies and against nonrelativistic CCSD/B2 for dipole moments.

while NDO methods yielded BMUEs between 1.964 (PM3, best) to 2.686 (MNDO, worst). The HF theory resulted in a BMUE of 1.222, less favorable than all but two of the studied DFT methods. With a BMUE of 1.964, PM3 is the best NDO method; however, PM3 still tested worse than all DFT methods we examined. While SCC-DFTB returned very inaccurate bond dissociation energies (see above) and was therefore not included in the calculations of BMUE, its performance was far better for bond lengths (MUE = 0.043 Å, best among the NDO methods) and somewhat better for dipole moments (MUE = 1.45 D, worst among the NDO methods).

The M05-2X functional,[58] which has demonstrated excellent performance for noncovalent interactions and barrier heights in tests against broad main-group databases[58,122] and tests for silicon chemistry,[123] gives the lowest overall nonrelativistic BMUE (0.333) for this compound set. M05-2X was parametrized against 34 nonmetal data values, whereas the closely related M05 functional was parametrized for metals as well as nonmetals.[57] Interestingly, here we find that M05-2X performs better for Zn than does M05 (nonrelativistic BMUE = 0.437), indicating that Zn, a d^{10} transition metal, may more closely resemble a main-group element than it does other transition metals. This may also explain the extremely poor performance of the MOHLYP functional, a transition-metal-parametrized GGSC method, with regard to geometries.

Table 14. Mean Unsigned Errors in Bond Length (Å) and Dipole Moment (D) with the 6-31+G(d,p) Basis for H, C, N, O, S[a]

bond length	MUE (Å)	dipole moment	MUE (D)
PBEh	0.0068	M05-2X	0.20
PW6B95	0.0068	B97-2	0.23
mPW1PW	0.0073	mPW1PW	0.25
M05-2X	0.0091	M05	0.25
B97-2	0.0094	MPW1B95	0.26
X3LYP	0.0096	PBEh	0.26
MPW1KCIS	0.0096	PW6B95	0.30
M06-L	0.0099	O3LYP	0.32
TPSS1KCIS	0.0102	OVWNS	0.34
TPSSh	0.0103	OPWL	0.34
M06	0.0110	TPSSh	0.34
B3LYP	0.0111	B1LYP	0.34
BP86	0.0112	TPSS1KCIS	0.35
TPSS	0.0118	M06-2X	0.37
B1LYP	0.0120	τ-HCTH	0.38
M05	0.0133	M06-L	0.39
τ-HCTH	0.0144	X3LYP	0.40
MPW1B95	0.0145	MOHLYP	0.40
O3LYP	0.0166	B3LYP	0.40
VSXC	0.0181	M06	0.41
M06-2X	0.0194	OV5LYP	0.41
G96LYP	0.0213	OLYP	0.45
MPWLYP1M	0.0218	VSXC	0.51
mPWLYP	0.0248	TPSSVWNS	0.53
HF	0.0248	TPSS	0.53

[a] Otherwise the same as Tables 7 and 9. Only the top 25 functionals are shown for each property.

Table 15. Balanced Mean Unsigned Errors (BMUEs, Unitless) for Three Databases of Zn−Ligand Compounds for DFT Methods (Relativistic)[a] and NDO Methods

method	BMUE	method	BMUE
M05-2X	0.414	X3LYP	0.527
B97-2	0.439	B3LYP	0.550
PW6B95	0.444	PM3	1.81
mPW1B95	0.457	AM1	1.84
M05	0.458	PM6	1.91
mPW1PW	0.476	MNDO	2.33
PBEh	0.500	MNDO(d)	2.35
B1LYP	0.504		

[a] Relativistic DFT/B2 and other methods tested against relativistic CCSD(T)/B2 for geometries and bond dissociation energies and against relativistic CCSD/B2 for dipole moments.

Also in the top five functionals tested were B97-2 (BMUE = 0.376), mPW1PW (BMUE = 0.426), M05 (BMUE = 0.437), and PW6B95 (BMUE = 0.438). The popular B3LYP theory ranks #10 overall (BMUE = 0.510), testing quite well for geometries (MUE = 0.008 Å) but less so for dipole moments (MUE = 0.44 D) and bond dissociation energies (MUE = 5.41 kcal/mol).

Inclusion of HF exchange is found to be very helpful for the Zn compounds we included in the present tests: all functionals with HF exchange resulted in nonrelativistic BMUEs below the normalized mean of 1.0, whereas 45% of the functionals with no HF exchange yielded nonrelativistic BMUEs above 1.0. Ten additional functionals with X

= 0 give BMUEs above 0.7: TPSS, OLYP, OV5LYP, VSXC, BP86, TPSSVWN5, OVWN5, OPWL, G96HLYP, and G96VWN5. The best theory level with no HF exchange is M06-L, a new local functional which recently performed best for a broad combination of main-group thermochemistry, thermochemical kinetics, organometallic/inorganometallic, and noncovalent interactions as well as for geometric parameters and vibrational frequencies.[59] Because local functions are considerably less expensive than nonlocal functionals for large systems, the M06-L local functional has been suggested[59] for calculations involving medium-to-large systems and/or simulations involving longer time scales.

Incorporating the kinetic energy density τ_σ tended to lower the BMUE. All functionals that include τ_σ resulted in BMUEs better than the mean, and three of the top five functionals include τ_σ: M05-2X and M05 in both exchange and correlation and PW6B95 in correlation. However, including τ_σ does not appear to be a requirement for a good Zn functional; the other two of the top five theory levels do not incorporate it. In general, HMGGA and MGGA methods tested favorably for Zn compounds; GGA, GGE, and GGSC methods were less suitable, and the one LSDA method we evaluated, SVWN5, was the least favorable functional.

For bond lengths and dipole moments, we also carried out complete tests of the same density functionals with the smaller 6-311+G(d,p)[124] basis set for H, C, N, O, and S combined with the Zn basis as used in B1. We found similar trends to the results presented here, and so these results are not presented in detail. However, it is useful to summarize them, and this is done in Table 14, which presents the 25 best methods for geometries and the 25 best methods for dipole moments. For geometries, some methods actually perform better with the smaller basis set, but on average the errors increase by approximately 10%. In contrast the errors in the dipole moments tend to decrease about 10% with the smaller basis. This either indicates that the smaller basis is better balanced or is an encouraging indication that one can achieve similar accuracy with augmented polarized double-ζ basis sets to what one can obtain with augmented polarized triple-ζ basis sets. (The performance of augmented polarized double-ζ basis sets relative to larger basis sets in DFT calculations is also discussed elsewhere.[100])

4.2. Relativistic Tests. One of the key findings of the present study is that scalar relativistic effects are not negligible for Zn compounds. Table 1 shows that when relativistic effects are added, Zn bond lengths all decrease, with an average change of 0.011 Å, which is a factor of 1.7 larger than the smallest mean unsigned error in Table 7. This decrease is as expected, since the direct relativistic effect[125] decreases the size of core *s* and *p* electrons. Table 3 shows that relativistic effects on dipole moments are less systematic, with six dipole moments decreasing by an average of 0.20 D and two increasing by an average of 0.16 D. The Zn−ligand bond energies are also sensitive to relativistic effects; seven of them increase by an average of 4.7 kcal/mol, and the other five decrease by an average of 2.1 kcal/mol. Table 11 shows that these average changes are larger than the mean unsigned errors in the six best functionals.

We considered relativistic effects for the ten density functionals that performed best in the tests (discussed in section 4.1) of nonrelativistic density functional calculations against nonrelativistic best estimates. Relativistic density functional calculations (that is, density functional calculations employing relativistic effective core potentials) were carried out for these ten functionals and were compared to the relativistic best-estimate results. These relativistic tests lead to the same conclusions as the nonrelativistic ones, that is, the functionals that perform well in the nonrelativistic tests also perform well in the relativistic ones. Table 15 shows the performance of DFT and NDO methods against relativistic benchmarks, where both DFT and CCSD(T) calculations incorporate the multielectron-fit (MEFIT,*R*) pseudopotential of Preuss et al. on Zn.[95] M05-2X remains the top functional, with B97-2, PW6B95, mPW1B95, and M05 in second through fifth places, in that order. The kinetic energy density τ_σ is present in four of these five best performing functionals, and the Hartree−Fock exchange ranges from 21 to 56%.

5. Summary and Concluding Remarks

We have presented nonrelativistic and relativistic databases of CCSD(T) geometric parameters and bond dissociation energies and CCSD dipole moments, for a set of Zn model compounds, and used them as benchmarks to test a variety of nonrelativistic and relativistic DFT methods and other molecular orbital methods. While the accuracy of the DFT methods we tested varies considerably, as measured by balanced mean unsigned error (BMUE), DFT overall significantly outperformed NDO (“semiempirical”) and tight-binding molecular orbital methods for our compound sets. Although NDO and tight-binding methods are parametrized against experimental data and therefore include electron correlation effects implicitly, it is disappointing that their overall errors (measured by BMUE against nonrelativistic benchmarks) are factors of 1.5−2.1 larger than ab initio Hartree−Fock. Two of the 38 density functionals we tested also fared worse in the nonrelativistic tests than ab initio Hartree−Fock. Seventeen density functionals (including five developed in Minnesota), however, have BMUEs more than a factor of 2 lower than HF.

Our results indicate that the suitability of a particular functional for Zn is enhanced by Hartree−Fock (HF) exchange and often, although not necessarily, by including the kinetic energy density τ_σ. The M05-2X functional has been recommended for general-purpose, nonmetal thermochemistry, kinetics, and noncovalent interactions.[58] Based on our analysis of nonrelativistic and relativistic mean unsigned errors (MUEs), in which M05-2X surpasses all other methods tested, we now recommend it to obtain accurate geometric parameters, dipole moments, and bond dissociation energies for Zn centers. For those interested in a broadly applicable local ($X = 0$) functional with reduced computational cost, perhaps to model larger Zn systems or for simulations of longer duration, we suggest the M06-L and τ-HCTH functionals, which display the best performance of the 20 local functionals studied.

Zinc represents an interesting borderline case in the periodic table:[126] it is sometimes considered a transition metal and sometimes a main-group element. The present study indicates that the computational chemistry requirements of Zn resemble those of the main group rather than the first transition row, as the best functionals for Zn are those that generally perform best for the main group.

Acknowledgment. The authors express their gratitude to Benjamin Lynch of the Minnesota Supercomputing Institute, Nate Schultz and Yan Zhao of the University of Minnesota Department of Chemistry, and Clayton F. Spencer of Illinois College for valuable assistance and discussions. This work was supported in part by the Office of Naval Research under award number N00014-05-1-0538 and by the National Science Foundation through grant no. CHE07-49122.

Supporting Information Available: Geometries at the X3LYP and M05-2X levels, dipole moments at the M05-2X and B97-2 levels, bond dissociation energies at the M05-2X and M05 levels, mean signed errors for all methods (relativistic/B2 and nonrelativistic/B1), and mean unsigned errors in bond length, dipole moment, and bond dissociation energy (relativistic/B2). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Auld, D. S. In *Handbook on Metalloproteins*; Bertini, I., Sigel, A., Sigel, H., Eds.; Dekker: New York, 2001; p 881.

(2) Dudev, T.; Lim, C. *Chem. Rev.* **2003**, *103*, 773.

(3) Silva, J. J. R. F. D.; Williams, R. J. P. *The Biological Chemistry of the Elements: The Inorganic Chemistry of Life*, 2nd ed.; Oxford University Press: New York, 2001.

(4) Wu, H. Z.; Qiu, D. J.; Cai, Y. J.; Xu, X. L.; Chen, W. B. *J. Cryst. Growth* **2002**, *245*, 50.

(5) Koca, J.; Zhan, C. G.; Rittenhouse, R. C.; Ornstein, R. L. *J. Comput. Chem.* **2003**, *24*, 368.

(6) Kikuchi, K.; Komatsu, K.; Nagano, T. *Curr. Opin. Chem. Biol*. **2004**, *8*, 182.

(7) Remko, M.; Garaj, V. *Mol. Phys.* **2003**, *101*, 2357.

(8) Porento, M.; Hirva, P. *Theor. Chem. Acc*. **2002**, *107*, 200.

(9) Dewar, M. J. S.; Merz, K. M., Jr. *Organometallics* **1988**, *7*, 522.

(10) Onciul, A. R. V.; Clark, T. *J. Comput. Chem*. **1993**, *14*, 392.

(11) Stanton, R. V.; Merz, K. M., Jr. *J. Chem. Phys*. **1994**, *100*, 434.

(12) Shoner, S. C.; Humphreys, K. J.; Barnhart, D.; Kovacs, J. A. *Inorg. Chem*. **1995**, *34*, 5933.

(13) Olson, L. P.; Luo, J.; Almarsson, O.; Bruice, T. C. *Biochemistry* **1996**, *35*, 9782.

(14) Gresh, N.; Garmer, D. R. *J. Comput. Chem*. **1996**, *17*, 1481.

(15) Vanhommeig, S. A. M.; Meier, R. J.; Stuyterman, L. A. A.; Meijer, E. M. *Theochem* **1996**, *364*, 33.

(16) Vaz, R. J.; Kuntz, I. D.; Meng, E. C. *Med. Chem. Res*. **1999**, *9*, 479.

(17) Amin, E. A.; Welsh, W. J. *J. Med. Chem*. **2001**, *44*, 3849.

(18) Yazal, J. E.; Pang, Y.-P. *Theochem* **2001**, *545*, 271.

(19) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.

(20) Remko, M.; Garaj, V. *Mol. Phys.* **2003**, *101*, 2357.

(21) Banci, L. *Curr. Opin. Chem. Biol*. **2003**, *7*, 143.

(22) Asthagiri, D.; Pratt, L. R.; Paulaitis, M. E.; Rempe, S. B. *J. Am. Chem. Soc*. **2004**, *125*, 1285.

(23) Linder, D. P.; Rodgers, K. R. *J. Phys. Chem. B* **2004**, *108*, 13839.

(24) Ambroggio, X. I.; Rees, D. C.; Deshales, R. J. *PLoS Biol.* **2004**, *2*, 113.

(25) Dudev, T.; Chang, L.-Y.; Lim, C. *J. Am. Chem. Soc*. **2005**, *127*, 4091.

(26) Martino, T.; Russo, N.; Toscano, M. *J. Am. Chem. Soc*. **2005**, *127*, 4242.

(27) Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921.

(28) Gresh, N.; Piquemal, J.-P.; Krauss, M. *J. Comput. Chem*. **2005**, *26*, 1113.

(29) Estiu, G.; Suarez, D.; Merz, K. M., Jr. *J. Comput. Chem*. **2006**, *27*, 1240.

(30) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Am. Chem. Soc*. **2007**, *129*, 1378.

(31) Yazal, J.E.; Pang, Y.-P. *J. Mol. Struct. Theochem* **2001,** *545*, 271.

(32) Kornhaber, G. J.; Snyder, D.; Moseley, H. N. B.; Montelione, G. T. *J. Biomol. NMR* **2006**, *34*, 259.

(33) Xu, D.; Guo, H.; Cui, Q. *J. Phys. Chem. A* **2007**, *111*, 5630.

(34) Fan, Y.; Gao, Y. Q. *J. Am. Chem. Soc.* **2007**, *129*, 905.

(35) Bergquist, C.; Fillebeen, T.; Morlok, M.; Parkin, G. *J. Am. Chem. Soc.* **2003**, *125*, 6189.

(36) Dudutiene, V.; Baranauskiene, L.; Matulis, D. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3335.

(37) Christianson, D. W.; Fierke, C. A. *Acc. Chem. Res.* **1996**, *29*, 331.

(38) Lavrov, E. V.; Weber, J.; Börrnert, F.; Van de Walle, C. G.; Helbig, R. *Phys. Rev. B* **2002**, *66*, 165205.

(39) Myong, S. Y.; Park, S. I.; Lim, K. S. *Thin Solid Films* **2006**, *513*, 148.

(40) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(41) Zerner, M. C. *Rev. Comp. Chem*. **1991**, *2*, 313.

(42) *Molecular Orbital Calculations for Biological Systems*; Sapse, A.-M., Ed.; Oxford University Press: New York, 1998.

(43) Rey, J.; Savin, A. *Int. J. Quantum Chem.* **1998**, *69*, 581.

(44) Krieger, J. B.; Chen, J.; Iafrate, G. J.; Savin, A. In *Electron Correlations and Materials Properties;* Gonis, A., Kioussis, N., Eds.; Plenum: New York, 1999; p 463.

(45) Toulouse, J.; Savin, A.; Adamo, C. *J. Chem. Phys.* **2002**, *117*, 10465.

(46) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett*. **2003**, *91*, 146401.

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

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

(49) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(50) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.

(51) Adamo, C.; Barone, V. *J. Chem. Phys*. **1998**, *108*, 664.

(52) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.

(53) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys*. **2003**, *119*, 12129; **2004**, *121*, 11507(E).

(54) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys*. **2001**, *115*, 9233.

(55) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.

(56) Gill, P. M. W. *Mol. Phys*. **1996**, *89*, 433.

(57) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(58) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

(59) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(60) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* In press.

(61) Slater, J. C. *Quantum Theory of Molecules and Solids*; McGraw-Hill: New York, 1974; Vol. 4.

(62) Vosko, S. H.; Wilk, L.; Nussair, M. *Can. J. Phys*. **1980**, *58*, 1200.

(63) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

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

(65) Perdew, J. P. In *Electronic Structure of Solids;* Ziesche, P., Eschrig, H., Eds.; Kademie Verlag: Berlin, 1991; p 11.

(66) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

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

(68) Stevens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem*. **1994**, *98*, 11623.

(69) Becke, A. D. *J. Chem. Phys*. **1996**, *104*, 1040.

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

(71) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242.

(72) Voorhis, T. V.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.

(73) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(74) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(75) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.

(76) Xu, X.; Goddard, W. A. *I. Proc. Natl. Acad. Sci*. **2004**, *101*, 2673.

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

(78) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

(79) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(80) Dewar, M. J. S.; Jie, C.; Yu, G. *Tetrahedron* **1993**, *23*, 5003.

(81) Holder, A. J.; Dennington, R. D.; Jie, C.; Yu, G. *Tetrahedron* **1994**, *50*, 627.

(82) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.

(83) Dewar, M. J. S.; Merz, K. M., Jr. *Organometallics* **1986**, *5*, 1494.

(84) Thiel, W.; Voityuk, A. *Theor. Chim. Acta* **1992**, *81*, 391.

(85) Thiel, W.; Voityuk, A. A. *J. Phys. Chem*. **1996**, *100*, 616.

(86) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(87) Hehre, W. J.; Yu, J.; Adei, E. *Abstracts of Papers*, 212th ACS National Meeting, Orlando, FL, Aug. 25−29, 1996; American Chemical Society: Washington, DC, 1996; COMP 092.

(88) Hehre, W. J.; Yu, J.; Adei, E. *A Guide of Molecular Mechanics and Molecular Orbital Calculations in SPARTAN*; Wavefunction Inc.: Irvine, CA, 1997.

(89) Cundari, T. R.; Deng, J. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 376.

(90) Bosque, R.; Maseras, F. *J. Comput. Chem*. **2000**, *21*, 562.

(91) Stewart, J. J. P. *Abstracts of Papers*, 232nd ACS National Meeting, San Francisco, CA, Sept. 10−14, 2006; American Chemical Society: Washington, DC, 2006; COMP 134.

(92) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.

(93) Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185.

(94) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

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

(96) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(97) Wachters, A. J. H. *J. Chem. Phys.* **1970**, *52*, 1033.

(98) Hay, P. J. *J. Chem. Phys*. **1977**, *68*, 4377.

(99) Raghavachari, K.; Trucks, G. W. *J. Chem. Phys*. **1989**, *91*, 1062.

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

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

(102) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(103) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

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

(105) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.

(106) Federov, D. G.; Koseki, S.; Schmidt, M. W.; Gordon, M. S. *Int. Rev. Phys. Chem.* **2003**, *22*, 551.

(107) Moore, C. National Bureau of Standard (U.S.) Circular 467, 2952.

(108) Schultz, N. E.; Gherman, B. F.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 24030.

(109) Shayesteh, A.; Yu, S.; Bernath, P. *Chem. Eur. J.* **2005, 11,** 4709.

(110) Shayesteh, A.; Gordon, I.; Appadoo, D.; Bernath, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3132.

(111) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.

(112) Bartlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1697.

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

(114) Kinal, A.; Piecuch, P. *J. Phys. Chem. A* **2007**, *111*, 734.

(115) Wong, B. M.; Raman, S. *J. Comput. Chem.* **2007**, *28*, 759.

(116) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(117) Hill, J. G.; Platts, J. A.; Werner, H. J. *Phys. Chem. Chem. Phys.* **2006**, *125*, 133206.

(118) Schultz, N. E.; Truhlar, D. G. Unpublished.

(119) Wavefunction, Inc., Irvine, CA.

(120) Stewart, J. J. P. http://mopac2007.net (accessed February 23, 2007).

(121) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 5121.

(122) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput*. **2007**, in press.

(123) Zhang, Y.; Li, Z. H.; Truhlar, D. G. *J. Chem. Theory Comput*. **2007**, ASAP.

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

(125) Rose, S. J.; Grant, I. P.; Pyper, N. J. *J. Phys. B* **1978**, *11*, 1171.

(126) Mann, J. B.; Meek, T. L.; Allen, L. C. *J. Am. Chem. Soc*. **2000**, *122*, 2780.

CT700205N

JCTC Journal of Chemical Theory and Computation

Comparison of the Standard 6-31G and Binning-Curtiss Basis Sets for Third Row Elements

Shahidul M. Islam, Stephanie D. Huelin, Margot Dawe, and Raymond A. Poirier*

Department of Chemistry, Memorial University,
St. John's, Newfoundland, Canada A1B 3X7

Received September 4, 2007

Abstract: Ab initio calculations were carried out for isogyric reactions involving the third row elements, Ga, Ge, As, Se, and Br. Geometries of all the reactants and products were optimized at the HF, MP2, and B3LYP levels of theory using the 6-31G(d) and 6-31G(d,p) basis sets. For molecules containing third row elements geometries, frequencies and thermodynamic properties were calculated using both the standard 6-31G and the Binning-Curtiss (BC6-31G) basis sets. In order to determine the performance of these basis sets, the calculated thermodynamic properties were compared to G3MP2 values and where possible to experimental values. Geometries and frequencies calculated with the standard 6-31G and the BC6-31G basis sets were found to differ significantly. Frequencies calculated with the standard 6-31G basis set were generally in better agreement with the experimental values (MAD=40.1 cm^{-1} at B3LYP/6-31G-(d,p) and 94.2 cm^{-1} at MP2/6-31G(d,p) for unscaled frequencies and 29.6 cm^{-1} and 24.4 cm^{-1}, respectively, for scaled frequencies). For all the reactions investigated, the thermodynamic properties calculated with the standard 6-31G basis set were found to consistently be in better agreement with the G3MP2 and the available experimental results. However, the BC6-31G basis set performs poorly for the reactions involving both second and third row elements. Since, in general, the standard 6-31G basis set performs well for all the reactions, we recommend that the standard 6-31G basis set be used for calculations involving third row elements. Using G3MP2 enthalpies of reaction and available experimental heats of formation (ΔH_f), previously unknown ΔH_f for CH_3SeH, SiH_3SeH, CH_3AsH_2, SiH_3AsH_2, CH_3GeH_3, and SiH_3GeH_3 were found to be 18.3, 18.0, 38.4, 82.4, 41.9, and 117.4 kJ mol^{-1}, respectively.

1. Introduction

Calculations for compounds containing first and second row elements are now very common. However, fewer calculations have been performed for compounds containing third row main group elements. Such computations require basis sets that are consistent with those used for the first and second row elements in order to obtain accurate and reliable results.[1] For third row elements the Binning-Curtiss (BC6-31G) basis set[2] has been used in combination with the standard 6-31G basis set in most electronic structure packages (for example, Gaussian[3] and GAMESS[4]). However, this basis set does not actually meet the definition of the standard 6-31G basis set, but it is constructed from a contraction of the Dunning basis set.[2] The core functions are highly contracted with respect to those which represent the valence region which are kept uncontracted in order to maintain flexibility. For example, for the BC6-31G basis set, the s, p, and d shells consist of six (821111), four (6311), and one (5) contracted functions, respectively, which gives a total of 24 basis functions. Rassolov et al.[5] have developed a standard 6-31G basis set for the third row elements to use in G3 theories,[6] where the 3d orbitals are included in the valence space of the third row elements, resulting in a total of 29 basis functions. However, very little has been reported[7,8] on the use of the

* Corresponding author phone: (709) 737-8609; fax: (709) 737-3702; e-mail: rpoirier@mun.ca.

10.1021/ct700224j CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/08/2007

standard 6-31G basis set vs the BC6-31G basis set for lower levels of theory (e.g., HF, MP2, and B3LYP) for compounds containing third row elements. No detailed investigation of geometries and frequencies for compounds containing third row elements have been performed using the standard 6-31G basis set. From our previous study,[7] thermodynamic properties for $SiH_3Br + HCN \rightarrow SiH_3CN + HBr$ and $SiHBr + H_2 \rightarrow SiH_2 + HBr$, calculated at HF, MP2, and B3LYP levels using the standard 6-31G basis set, were found to be in better agreement with Gaussian-n theories compared to values obtained using the BC6-31G basis set for bromine with the same levels of theory. However, in a later study[8] on the reaction of alkenes with bromine, the thermodynamic properties obtained with the standard 6-31G basis set were found to agree very well with the values obtained with the BC6-31G basis set. In this study, we have extended the study to encompass other third row main group elements. G3MP2 theory[9] and experimental results where available are used in this study in evaluating the performance of the standard 6-31G and the BC6-31G basis sets for compounds containing first row and second row elements in combination with the third row elements Ga, Ge, As, Se, and Br. The results are compared, contrasted, and evaluated at the HF, MP2, and B3LYP levels of theory.

2. Method

In this study, the performance of the third row basis sets is evaluated by comparing the thermodynamics properties for the following isogyric reactions:

$$CH_3XH_n + HCN \rightarrow CH_3CN + XH_{n+1} \qquad (1)$$

$$SiH_3XH_n + HCN \rightarrow SiH_3CN + XH_{n+1} \qquad (2)$$

where X = Ga, Ge, As, Se, and Br and n = 2, 3, 2, 1, and 0, respectively, and for

$$CH_3Br + HCl \rightarrow CH_3Cl + HBr \qquad (3)$$

$$SiH_3Br + HCl \rightarrow SiH_3Cl + HBr \qquad (4)$$

$$PH_2Br + HCN \rightarrow PH_2CN + HBr \qquad (5)$$

The geometries and frequencies for molecules containing third row atoms are also investigated. All the electronic structure calculations were carried out with Gaussian03.[3] The geometries of all reactants and products were fully optimized at the HF, MP2, and B3LYP levels of theory using both the 6-31G(d) and 6-31G(d,p) basis sets. For third row elements, Ga, Ge, As, Se, and Br, both the standard 6-31G[5] and BC6-31G[2] basis sets are used throughout. Geometries of all the compounds were optimized ensuring all had their expected symmetries. From our previous work,[7,8] it was found that the enthalpies of reaction calculated by using G3MP2,[9] G3B3,[10] and G3MP2B3[10] levels of theory all agreed to within 5.3 kJ mol^{-1}. Therefore, the G3MP2 level of theory is used in this study which is expected to adequately reproduce the experimental data. G3MP2 theory is based on geometry optimizations performed at the MP2(full) level of theory using the standard 6-31G(d) basis set for first, second, and third row elements. In some cases G3MP2 calculations were also performed using the BC6-31G(d) basis set for third row elements. For the standard 6-31G(d) basis set, the G3MP2 energy is the summation of the following single point energies

$$E[\text{QCISD(T)/6-31G(d)}] + \Delta E_{\text{MP2}} + \Delta E(\text{SO}) + E(\text{HLC}) + E(\text{ZPE})$$

where

$$(\Delta E_{\text{MP2}}) = [E(\text{MP2/G3MP2Large})] - [E(\text{MP2/6-31G(d)})]$$

While the G3MP2 energy calculated using the BC6-31G(d) basis set is given by

$$E[\text{QCISD(T)/BC6-31G(d)}] + \Delta E_{\text{MP2}} + \Delta E(\text{SO}) + E(\text{HLC}) + E(\text{ZPE})$$

where

$$(\Delta E_{\text{MP2}}) = [E(\text{MP2/G3MP2Large})] - [E(\text{MP2/BC6-31G(d)})]$$

For all the third row elements the G3MP2Large basis set,[1,11] which is not yet incorporated in Gaussian03, was used for G3MP2 calculations. Frequencies were calculated for all structures to ensure the absence of imaginary frequencies in the minima.

3. Results and Discussion

The optimized geometries, frequencies, the thermodynamic properties of the isogyric reactions, and heats of formation of some energetically stable compounds containing third row elements are presented in Tables 1−9.

3.1. Geometries of Molecules Containing Third Row Elements. Bond lengths and angles calculated at the MP2 and B3LYP levels of theory using the standard 6-31G(d,p) and BC6-31G(d,p) basis sets for all the structures containing third row elements are listed in Table 1 along with the experimental data where available.

The geometric parameters calculated with the standard 6-31G(d,p) and BC6-31G(d,p) basis sets are quite different. The MP2 bond lengths are always shorter than the B3LYP bond lengths, and with a few exceptions (∠H−C−H in CH_3Br, CH_3SeH, CH_3AsH_2, and CH_3GaH_2 and ∠H−Si−X (X=Br, Ga) in SiH_3Br and SiH_3GaH_2) the MP2 bond angles are larger or almost equal to B3LYP angles. However, for all the levels of theory the agreement with experiment is similar to that found for compounds containing first and second row elements.

Table 2 lists the mean absolute deviations (MAD) in bond lengths and angles from experiment and calculations. A total of 25 experimental bond lengths and 18 experimental bond angles were used to calculate the mean absolute deviations from experiment. A total of 36 bond lengths and 36 bond angles were used to calculate the mean absolute deviations between the values calculated at the MP2 and the B3LYP level of theory using the standard 6-31G(d,p) and BC6-31G(d,p) basis sets. For bond lengths the MAD is ~0.012 Å except for B3LYP/6-31G(d,p) which has a MAD of 0.019 Å. The lowest MAD (0.0118 Å) is given by MP2/6-31G(d).

Table 1. Optimized and Experimental Structural Parameters for Compounds Containing Third Row Elements[t]

			MP2			B3LYP			
molecules	point group	geometric parameter	/6-31G(d,p)	/BC6-31G(d,p)	Δ^s	/6-31G(d,p)	/BC6-31G(d,p)	Δ^s	exptl
HBr	$C_{\infty v}$	H−Br	1.4075	1.4057	0.0018	1.4269	1.4171	0.0098	1.4144,[a] 1.4129[b]
SeH_2	C_{2v}	Se−H	1.4527	1.4480	0.0047	1.4738	1.4614	0.0124	1.4600,[a] 1.4605[b]
		∠H−Se−H	91.6	91.5	0.1	91.2	91.0	0.2	90[c]
AsH_3	C_{3v}	As−H	1.5042	1.5043	−0.0001	1.5271	1.5181	0.009	1.5108,[a] 1.5187[b]
		∠H−As−H	93.0	92.2	0.8	91.9	91.2	0.7	90[c]
GeH_4	T_d	Ge−H	1.5219	1.5285	−0.0066	1.5369	1.5306	0.0063	1.5151,[a] 1.5293,[b] 1.514[d]
		∠H−Ge−H	109.5	109.5	0.0	109.5	109.5	0.0	109.5[c]
GaH_3	D_{3h}	Ga−H	1.5579	1.5785	−0.0206	1.5700	1.5733	−0.0033	1.560,[a] 1.5505[e]
		∠H−Ga−H	120.0	120.0	0.0	120.0	120.0	0.0	
PH_2Br	*Cs*	P−Br	2.2474	2.2440	0.0034	2.2775	2.2612	0.0163	2.234,[f] 2.230[f]
		P−H	1.4067	1.4063	0.0004	1.4248	1.4242	0.0006	1.425,[f] 1.412[f]
		Br−H	2.7894	2.7847	0.0047	2.8183	2.8106	0.0077	
		∠H−P−Br	96.8	96.7	0.1	96.4	96.8	−0.4	96.1[f]
		∠H−P−H	93.5	93.4	0.1	92.2	92.1	0.1	92.4[f]
SiHBr	*Cs*	Si−Br	2.2529	2.2470	0.0059	2.2809	2.2601	0.0208	2.237,[g] 2.231[h]
		Si−H	1.5086	1.5082	0.0004	1.5308	1.5309	−0.0001	1.518,[g] 1.561[h]
		∠H−Si−Br	94.5	94.6	−0.1	94.2	94.6	−0.4	93.4[g]
CH_3Br	C_{3v}	C−Br	1.9424	1.9480	−0.0056	1.9658	1.9625	0.0033	1.939,[i] 1.934,[j] 1.933[k]
		C−H	1.0832	1.0834	−0.0002	1.0879	1.0878	0.0001	1.113,[i] 1.082,[j] 1.086[k]
		∠H−C−Br	108.1	107.8	0.3	107.7	107.7	0.0	107.7[j]
		∠H−C−H	110.8	111.1	−0.3	111.2	111.2	0.0	111.2,[j] 111.17[k]
SiH_3Br	C_{3v}	Si−Br	2.2294	2.2249	0.0045	2.2484	2.2299	0.0185	2.212,[l] 2.210[m]
		Si−H	1.4690	1.4686	0.0004	1.4808	1.4808	0.0	1.474,[l] 1.487[m]
		∠H−Si−Br	108.4	108.4	0.0	108.5	108.7	−0.2	108.2[l]
		∠H−Si−H	110.5	110.5	0.0	110.4	110.2	0.2	
CH_3SeH	*Cs*	C−Se	1.9610	1.9503	0.0107	1.9812	1.9633	0.0179	1.976[n]
		C−H	1.0896	1.0899	−0.0003	1.0909	1.0912	−0.0003	1.10[n]
		Se−H	1.4730	1.4799	−0.0069	1.4827	1.4848	−0.0021	1.48[n]
		∠H−C−H	110.7	110.9	−0.2	110.9	110.9	0.0	111[n]
		∠C−Se−H	95.0	95.8	−0.8	94.9	95.6	−0.7	95[n]
SiH_3SeH	*Cs*	Si−Se	2.2909	2.2895	0.0014	2.3086	2.2963	0.0123	
		Si−H	1.4816	1.4810	0.0006	1.4851	1.4846	0.0005	
		Se−H	1.4741	1.4812	−0.0071	1.4829	1.4849	−0.0020	
		∠H−Si−H	110.2	110.3	−0.1	109.9	109.9	0.0	
		∠Si−Se−H	93.9	94.5	−0.6	93.6	94.2	−0.6	
CH_3AsH_2	*Cs*	C−As	1.9798	1.9607	0.0191	1.999	1.983	0.016	1.92[o]
		C−H	1.0924	1.0928	−0.0004	1.0920	1.0916	0.0004	1.09[o]
		As−H	1.5248	1.5354	−0.0106	1.5300	1.5205	0.0095	
		∠H−C−H	109.4	109.2	0.2	109.6	109.9	−0.3	
		∠C−As−H	96.0	96.5	−0.5	95.6	95.1	0.5	
SiH_3AsH_2	*Cs*	Si−As	2.3705	2.3672	0.0033	2.3949	2.3698	0.0251	
		Si−H	1.4838	1.4836	0.0002	1.4873	1.4857	0.0016	
		As−H	1.5243	1.5357	−0.0114	1.5347	1.5186	0.0161	1.52[o]
		∠H−Si−H	109.1	109.2	−0.1	108.8	108.8	0.0	109.28[o]
		∠Si−As−H	93.6	93.8	−0.2	92.8	93.5	−0.7	94[o]
CH_3GeH_3	C_{3v}	C−Ge	1.9540	1.9474	0.0066	1.9692	1.9515	0.0177	1.9490,[p] 1.9453[q]
		C−H	1.0873	1.0874	−0.0001	1.0924	1.0924	0.0	1.0921,[p] 1.083[q]
		Ge−H	1.5264	1.5324	−0.0060	1.5414	1.5361	0.0053	1.5285,[p] 1.529[q]
		∠H−C−H	108.7	108.8	−0.1	108.7	108.7	0.0	108.841,[p] 108.4[q]
		∠C−Ge−H	110.3	110.6	−0.3	110.2	110.6	−0.4	109.3[q]
		∠H−Ge−H	108.5	108.3	0.2	108.4	108.3	0.1	108.776[p]
SiH_3GeH_3	C_{3v}	Si−Ge	2.3838	2.3828	0.0010	2.3987	2.3795	0.0192	2.36[r]
		Si−H	1.4761	1.4758	0.0003	1.4872	1.4868	0.0004	1.49[r]
		Ge−H	1.5252	1.5337	−0.0085	1.5400	1.5372	0.0028	1.53[r]
		∠Si−Ge−H	110.7	110.7	0.0	110.8	110.8	0.0	
		∠H−Si−H	108.8	108.9	−0.1	108.6	108.6	0.0	108.8[r]
		∠H−Ge−H	108.2	108.2	0.0	108.1	108.1	0.0	108.8[r]
CH_3GaH_2	*Cs*	C−Ga	1.9686	1.9874	−0.0188	1.9796	1.9771	0.0025	
		Ga−H	1.5636	1.5840	−0.0204	1.5769	1.5800	−0.0031	

Table 1. (Continued)

molecules	point group	geometric parameter	MP2 /6-31G(d,p)	MP2 /BC6-31G(d,p)	MP2 Δ^{s}	B3LYP /6-31G(d,p)	B3LYP /BC6-31G(d,p)	B3LYP Δ^{s}	exptl
		C−H	1.0919	1.0928	−0.0009	1.0974	1.0980	−0.0006	
		∠C−Ga−H	120.6	120.9	−0.3	120.6	121.0	−0.4	
		∠H−Ga−H	118.8	118.2	0.6	118.7	118.0	0.7	
		∠Ga−C−H	108.6	109.1	−0.5	108.6	108.9	−0.3	
		∠Ga−C−H	111.9	112.2	−0.3	111.8	112.1	−0.3	
		∠H−C−H	107.5	107.1	0.4	107.5	107.2	0.3	
		∠H−C−H	109.3	108.9	0.4	109.5	109.1	0.4	
SiH_3GaH_2	*Cs*	Si−Ga	2.4212	2.4199	0.0013	2.4315	2.4004	0.0311	
		Ga−H	1.5626	1.5830	−0.0204	1.5762	1.5789	−0.0027	
		Si−H	1.4808	1.4816	−0.0008	1.4922	1.4929	−0.0007	
		∠Si−Ga−H	121.0	121.5	−0.5	121.0	121.6	−0.6	
		∠H−Ga−H	118.1	116.9	1.2	117.9	116.7	1.2	
		∠Ga−Si−H	108.8	108.8	0.0	108.7	108.6	0.1	
		∠Ga−Si−H	112.0	112.1	−0.1	112.4	112.6	−0.2	
		∠H−Si−H	107.8	107.7	0.1	107.5	107.4	0.1	
		∠H−Si−H	108.3	108.3	0.0	108.2	108.1	0.1	

[a] Reference 14. [b] Reference 15. [c] Reference 16. [d] Reference 17. [e] Reference 18. [f]Reference 19. [g] Reference 20. [h] Reference 21. [i] Reference 22. [j] Reference 23. [k] Reference 24. [l] Reference 25. [m] Reference 26. [n] Reference 27. [o] Reference 28. [p] Reference 29. [q] Reference 30. [r] Reference 31. [s] Δ represents the difference between parameters calculated with the standard 6-31G(d,p) and the BC6-31G(d,p) basis sets. [t] Bond lengths are in Å and angles are in deg.

Table 2. Mean Absolute Deviations for Bond Lengths and Angles[a]

comparison	MAD (bond lengths)	MAD (angles)
experiment vs		
MP2/6-31G(d,p)	0.0118	0.6
MP2/BC6-31G(d,p)	0.0124	0.6
B3LYP/6-31G(d,p)	0.0187	0.5
B3LYP/BC6-31G(d,p)	0.0124	0.5
MP2/6-31G(d,p) vs		
MP2/BC6-31G(d,p)	0.0057	0.3
B3LYP/BC6-31G(d,p) vs		
B3LYP/6-31G(d,p)	0.0078	0.3

[a] Mean absolute deviations from experiment were calculated from 25 bond lengths and 18 bond angles, while 36 bond lengths and 36 bond angles were used to calculate the MAD between the calculated bond lengths and angles. Bond lengths are in Å, and angles are in deg.

For bond angles the MAD (18 of bond angles) is 0.5−0.6. Changes in bond lengths with a change in basis set are generally larger at B3LYP (0.0078 Å) than at MP2 (MAD=0.0057 Å). For example in HBr, the difference in bond lengths calculated at B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p) is 0.0098 Å, while the difference at MP2/6-31G(d,p) and MP2/BC6-31G(d,p) is 0.0018 Å. However, the difference due to a change of basis set at the MP2 level for X−H bond distances in GaH_3, CH_3SeH, SiH_3SeH, SiH_3GeH_3, CH_3GaG_2, and SiH_3GaH_2 and C−X bond distances in CH_3Br and CH_3AsH_2 are larger than the respective B3LYP values.

3.2. Frequencies of Molecules Containing Third Row Elements. Frequencies for the molecules containing third row elements at MP2/6-31G(d,p), MP2/BC6-31G(d,p), B3LYP/6-31G(d,p), and B3LYP/BC6-31G(d,p) are listed in Table 3 along with the experimental frequencies where available. MAD values for the frequencies are given in Table 4. A total of 145 frequencies of compounds containing third row elements were used to calculate the MAD between calculated frequencies and 73 to calculate the MAD between experimental and calculated frequencies. In most cases the B3LYP/6-31G(d,p) frequencies are in better agreement with experimental frequencies (Tables 3 and 4), with a MAD of 40.1 cm^{-1} compared to 57.8 cm^{-1} for B3LYP/BC6-31G(d,p) and 94.2 cm^{-1} and 105.4 cm^{-1} for MP2/6-31G(d,p) and MP2/BC6-31G(d,p), respectively. Therefore, for both MP2 and B3LYP the standard 6-31G basis set gives the best agreement, and overall the B3LYP with the standard 6-31G(d,p) basis set performs the best in calculating frequencies for molecules containing third row elements. B3LYP frequencies are found to be slightly more sensitive to the basis set than MP2 frequencies, i.e., the differences between the frequencies calculated at B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p), $\Delta\nu$(B3LYP), are generally larger than the differences between the frequencies calculated at MP2/6-31G(d,p) and MP2/BC6-31G(d,p), $\Delta\nu$(MP2) (Table 3). For unscaled frequencies the MAD between MP2/6-31G(d,p) and MP2/BC6-31G(d,p) is 16.4 cm^{-1}, while between B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p) the MAD is 20.3 cm^{-1}. The MAD between the MP2/6-31G(d,p) and B3LYP/6-31G(d,p) is 58.9 cm^{-1}, while the MAD between the MP2/BC6-31G(d,p) and B3LYP/BC6-31G(d,p) is 51.8 cm^{-1}, when unscaled frequencies are used. Standard frequency scaling factors for compounds containing first and second row elements are available in the literature.[12,13] The MAD for scaled frequencies using the standard scale factors are also given in Table 4. Scaling improves the frequencies significantly at all levels of theory and basis sets. After scaling MP2/6-31G(d,p) now has the lowest MAD (24.4 cm^{-1}) from experiment. For B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p) the MAD are lowered to 29.6 and 29.3 kJ mol^{-1}, respectively, when frequencies are scaled. The MAD between B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p) is 19.5 cm^{-1}, while MP2/6-31G(d,p) and MP2/BC6-31G(d,p) is 15.3 cm^{-1}.

Table 3. Calculated and Experimental Frequencies (in cm^{-1}) for Compounds Containing Third Row Elements[a]

			MP2			B3LYP			
molecules	point group	freq	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^m$	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^m$	expt
HBr	$C_{\infty v}$	ν_1	2759.3	2765.6	−6.3	2622.9	2663.3	−40.4	2558.5[b]
SeH_2	C_{2v}	ν_1	1125.1	1162.3	−37.2	1074.5	1131.5	−57.0	1034.2[c]
		ν_2	2544.3	2595.3	−51.0	2395.5	2449.6	−54.1	2344.5[c]
		ν_3	2564.0	2611.8	−47.8	2412.3	2466.2	−53.9	2357.8[c]
AsH_3	C_{3v}	ν_1	980.8	986.7	−5.9	946.3	968.7	−22.4	906.0[c]
		ν_2(e)	1079.2	1116.0	−36.8	1031.1	1071.7	−40.6	1003[c]
		ν_3	2315.1	2380.0	−64.9	2182.4	2261.6	−79.2	2116.1[c]
		ν_4(e)	2332.4	2395.0	−62.6	2200.7	2282.3	−81.6	2123.0[c]
GeH_4	T_d	ν_1(t_2)	861.3	851.3	10.0	823.6	820.1	3.5	819[d]
		ν_2(e)	965.1	956.3	8.8	928.9	935.3	−6.4	931[d]
		ν_3	2245.8	2332.0	−86.2	2138.3	2252.0	−113.7	2114[d]
		ν_4(t_2)	2247.4	2340.1	−92.7	2148.7	2273.6	−124.9	
GaH_3	D_{3h}	ν_1	750.4	730.1	20.3	718.4	711.3	7.1	717.4[e,f]
		ν_2(e)	792.5	784.3	8.2	762.5	776.5	−14.0	758.7[e,g]
		ν_3(e)	2033.8	2039.8	−6.0	1966.6	2018.2	−51.6	1923.2[e,g]
		ν_4	2038.2	2049.9	−11.7	1961.3	2012.2	−50.9	
PH_2Br	Cs	ν_1	412.8	423.9	−11.1	383.0	398.7	−15.7	399.79[h]
		ν_2	818.7	821.4	−2.7	784.9	794.2	−9.3	794.90[h]
		ν_3	863.9	869.8	−5.9	819.1	831.1	−12.0	812.46[h]
		ν_4	1165.6	1165.8	−0.2	1135.5	1138.1	−2.6	
		ν_5	2524.0	2522.6	1.4	2389.0	2387.2	1.8	
		ν_6	2537.7	2537.2	0.5	2401.9	2400.9	1.0	
$SiHBr$	Cs	ν_1	422.7	432.4	−9.6	394.9	410.1	−15.2	424.3[i]
		ν_2	815.7	820.7	−5.0	774.7	785.3	−10.6	553.6[i]
		ν_3	2164.4	2164.8	−0.4	2039.7	2038.1	1.6	1970.9[i]
CH_3Br	C_{3v}	ν_1	639.0	632.0	7.0	588.4	592.8	−4.4	617,[j] 611[k,c]
		ν_2(e)	1009.3	1003.9	5.4	968.1	967.6	0.5	974[j]
		ν_3	1405.2	1394.9	10.3	1345.8	1343.2	2.6	1333[j]
		ν_4(e)	1536.5	1540.4	−3.9	1487.7	1490.5	−2.8	1472[j]
		ν_5	3177.4	3173.9	3.5	3097.2	3096.1	1.1	3082,[j] 2972[k]
		ν_6(e)	3304.2	3302.5	1.7	3211.4	3210.0	1.4	3184[j]
SiH_3Br	C_{3v}	ν_1	441.9	448.1	−6.2	414.9	429.2	−14.3	430[c]
		ν_2(e)	655.4	668.4	−13	628.9	643.8	−14.9	633[c]
		ν_3	991.4	1000.1	−8.7	944.5	957.9	−13.4	930[c]
		ν_4(e)	999.0	1001.4	−2.4	954.9	955.2	−0.3	950[c]
		ν_5	2356.2	2360.1	−3.9	2253.3	2254.7	−1.4	2200[c]
		ν_6(e)	2374.5	2378.1	−3.6	2271.7	2271.8	−0.1	2196[c]
CH_3SeH	Cs	ν_1	212.0	229.3	−17.3	198.3	181.2	17.1	145[l]
		ν_2	614.0	606.1	7.9	572.8	571.1	1.7	584[l]
		ν_3	744.9	764.2	−19.3	715.1	744.3	−29.2	712[l]
		ν_4	961.1	950.2	10.9	919.5	914.7	4.8	921[l]
		ν_5	1046.8	1052.0	−5.2	1009.7	1022.8	−13.1	980[l]
		ν_6	1386.1	1379.2	6.9	1329.8	1329.7	0.1	1288[l]
		ν_7	1530.0	1536.9	−6.9	1485.1	1490.4	−5.3	1433[l]
		ν_8	1543.0	1548.0	−5.0	1494.6	1498.5	−3.9	1447[l]
		ν_9	2535.1	2582.3	−47.2	2378.9	2425.4	−46.5	2330[l]
		ν_{10}	3163.0	3162.5	0.5	3083.6	3082.4	1.2	2955[l]
		ν_{11}	3277.1	3277.2	−0.1	3182.4	3181.0	1.4	3027[l]
		ν_{12}	3284.0	3286.7	−2.7	3190.6	3190.6	0.0	3032[l]
SiH_3SeH	Cs	ν_1	184.4	175.2	9.2	175.9	102.1	73.8	
		ν_2	412.3	420.6	−8.3	386.7	399.8	−13.1	
		ν_3	529.2	554.5	−25.3	507.6	537.4	−29.8	
		ν_4	626.4	641.3	−14.9	598.2	615.0	−16.8	
		ν_5	779.1	802.4	−23.3	754.6	784.4	−29.8	
		ν_6	972.9	981.9	−9.0	924.4	937.2	−12.8	
		ν_7	982.0	983.2	−1.2	939.5	938.2	1.3	
		ν_8	1014.8	1020.1	−5.3	970.9	976.5	−5.6	
		ν_9	2336.4	2337.6	−1.2	2235.1	2235.3	−0.2	
		ν_{10}	2346.3	2347.0	−0.7	2244.5	2243.6	0.9	

Table 3. (Continued)

			MP2			B3LYP			
molecules	point group	freq	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^m$	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^m$	exptl
		ν_{11}	2361.3	2365.0	−3.7	2260.4	2262.6	−2.2	
		ν_{12}	2527.4	2561.9	−34.5	2378.8	2409.5	−30.7	
CH_3AsH_2	*Cs*	ν_1	206.2	238.4	−32.2	195.4	224.1	−28.7	
		ν_2	589.7	590.0	−0.3	554.1	555.5	−1.4	
		ν_3	667.2	699.3	−32.1	651.3	680.6	−29.3	
		ν_4	703.5	726.0	−22.5	678.6	701.7	−23.1	
		ν_5	966.8	959.3	7.5	932.3	930.3	2.0	
		ν_6	999.6	1015.4	−15.8	964.8	985.2	−20.4	
		ν_7	1057.8	1097.6	−39.8	1009.9	1050.9	−41.0	
		ν_8	1356.8	1351.4	5.4	1305.4	1305.9	−0.5	
		ν_9	1530.6	1542.1	−11.5	1488.4	1497.6	−9.2	
		ν_{10}	1534.9	1544.3	−9.4	1490.2	1498.4	−8.2	
		ν_{11}	2297.1	2365.1	−68.0	2157.8	2227.9	−70.1	
		ν_{12}	2309.7	2374.2	−64.5	2172.4	2241.9	−69.5	
		ν_{13}	3152.3	3154.0	−1.7	3070.9	3071.2	−0.3	
		ν_{12}	3259.2	3262.0	−2.8	3157.7	3160.3	−2.6	
		ν_{13}	3272.0	3278.0	−6.0	3178.2	3179.4	−1.2	
SiH_3AsH_2	*Cs*	ν_1	162.4	143.9	18.5	135.7	128.6	7.1	
		ν_2	376.6	379.7	−3.1	350.0	357.1	−7.1	
		ν_3	462.5	500.8	−38.3	444.1	485.8	−41.7	
		ν_4	481.2	515.3	−34.1	458.9	497.9	−39	
		ν_5	704.6	748.5	−43.9	681.4	732.1	−50.7	
		ν_6	758.9	805.7	−46.8	726.0	778.3	−52.3	
		ν_7	950.3	956.1	−5.8	902.7	912.6	−9.9	
		ν_8	990.0	992.7	−2.7	950.0	951.6	−1.6	
		ν_9	1000.3	1002.2	−1.9	958.4	959.8	−1.4	
		ν_{10}	1046.4	1084.8	−38.4	996.8	1046.5	−49.7	
		ν_{11}	2298.1	2324.2	−26.1	2173.2	2225.7	−52.5	
		ν_{12}	2312.1	2338.8	−26.7	2187.5	2238.3	−50.8	
		ν_{13}	2322.5	2341.8	−19.3	2223.6	2240.6	−17.0	
		ν_{12}	2337.6	2368.8	−31.2	2239.1	2250.8	−11.7	
		ν_{13}	2341.3	2376.5	−35.2	2241.1	2261.6	−20.5	
CH_3GeH_3	C_{3v}	ν_1	177.9	193.0	−15.1	158.3	183.0	−24.7	157^c
		ν_2(e)	506.9	496.3	10.6	493.4	490.1	3.3	506^c
		ν_3	616.0	637.3	−21.3	586.0	613.9	−27.9	$602^{c,m}$
		ν_4	882.4	872.4	10.0	848.1	845.4	2.7	$843^{c,m}$
		ν_5(e)	886.7	876.8	9.9	857.4	857.5	−0.1	848^c
		ν_6(e)	942.0	934.0	8.0	905.1	913.7	−8.6	900^c
		ν_7	1340.7	1332.6	8.1	1297.0	1295.4	1.6	$1254^{c,m}$
		ν_8(e)	1525.5	1528.4	−2.9	1484.0	1486.7	−2.7	1428^c
		ν_9(e)	2222.4	2312.4	−90	2126.0	2244.6	−118.6	$2085^{c,m}$
		ν_{10}	2223.9	2317.3	−93.4	2129.3	2259.2	−129.9	2084^c
		ν_{11}	3147.0	3147.2	−0.2	3063.8	3065.2	−1.4	$2938^{c,m}$
		ν_{12}(e)	3255.7	3257.0	−1.3	3153.4	3154.9	−1.5	2997^c
SiH_3GeH_3	C_{3v}	ν_1	122.2	122.3	−0.1	109.2	131.5	−22.3	144^n
		ν_2	370.1	356.6	13.5	348.4	340.1	8.3	$312,^n 318^m$
		ν_3(e)	376.7	369.5	7.2	370.2	371.0	−0.8	371^n
		ν_4(e)	627.1	619.5	7.6	600.1	602.5	−2.4	550^n
		ν_5	825.3	818.1	7.2	794.2	796.2	−2.0	$780,^n 785.2^m$
		ν_6(e)	926.6	916.0	10.6	889.3	899.1	−9.8	881^n
		ν_7	948.9	943.9	5.0	904.6	905.9	−1.3	$890,^n 890.3^m$
		ν_8(e)	997.6	997.2	0.4	955.5	955.4	0.1	930^n
		ν_9	2218.7	2294.8	−76.1	2124.5	2221.3	−96.8	$2052,^n 2076.6^m$
		ν_{10}(e)	2223.9	2305.7	−81.8	2134.1	2235.2	−101.1	2069^n
		ν_{11}	2319.3	2319.2	0.1	2222.5	2235.7	−13.2	$2151,^n 2163.1^m$
		ν_{12}(e)	2334.0	2334.5	−0.5	2236.9	2254.8	−17.9	2160^n
CH_3GaH_2	*Cs*	ν_1	10.5	36.4	−25.9	37.3	30.2	7.1	
		ν_2	430.2	417.5	12.7	418.2	419.2	−1.0	
		ν_3	519.2	514.1	5.1	501.8	498.7	3.1	

Table 3. (Continued)

			MP2			B3LYP			
molecules	point group	freq	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^{m}$	/6-31G(d,p)	/BC6-31G(d,p)	$\Delta\nu^{m}$	exptl
		ν_4	586.4	597.3	−10.9	560.0	578.9	−18.9	
		ν_5	769.1	758.1	11.0	750.4	748.9	1.5	
		ν_6	805.2	792.5	12.7	773.9	781.2	−7.3	
		ν_7	821.1	806.0	15.1	800.0	792.4	7.6	
		ν_8	1299.5	1293.8	5.7	1256.1	1254.0	2.1	
		ν_9	1510.7	1510.3	0.4	1470.2	1468.6	1.6	
		ν_{10}	1520.7	1518.9	1.8	1477.0	1475.4	1.6	
		ν_{11}	2004.7	2019.9	−15.2	1933.4	1993.3	−59.9	1892.0[g]
		ν_{12}	2012.1	2030.1	−18.0	1935.6	1994.3	−58.7	1898.0[g]
		ν_{13}	3127.6	3123.8	3.8	3039.5	3038.6	0.9	
		ν_{14}	3221.7	3214.6	7.1	3115.0	3111.7	3.3	
		ν_{15}	3253.5	3245.4	8.1	3148.2	3144.9	3.3	
SiH_3GaH_2	*Cs*	ν_1	6.6	13.0	−6.4	29.5	35.6	−6.1	
		ν_2	336.7	326.1	10.6	318.6	323.9	−5.3	
		ν_3	339.8	334.8	5.0	325.9	332.5	−6.6	
		ν_4	407.3	394.2	13.1	379.8	381.0	−1.2	
		ν_5	573.2	562.7	10.5	544.7	549.4	−4.7	
		ν_6	618.5	604.9	13.6	591.3	591.5	−0.2	
		ν_7	781.7	772.5	9.2	753.0	769.5	−16.5	
		ν_8	933.1	929.5	3.6	884.3	888.8	−4.5	
		ν_9	991.1	990.0	1.1	947.7	946.1	1.6	
		ν_{10}	999.3	998.7	0.6	956.5	956.1	0.4	
		ν_{11}	2006.3	2023.9	−17.6	1932.6	1994.8	−62.2	
		ν_{12}	2009.2	2026.2	−17	1940.7	2001.5	−60.8	
		ν_{13}	2297.2	2293.0	4.2	2198.4	2195.5	2.9	
		ν_{14}	2314.2	2311.0	3.2	2216.5	2215.2	1.3	
		ν_{15}	2321.0	2318.3	2.7	2224.2	2222.9	1.3	

[a] Calculated frequencies are not scaled. [b] Reference 32. [c] Reference 33. [d] Reference 34. [e] Reference 35. [f] Reference 36. [g] Reference 37. [h] Reference 38. [i] Reference 20. [j] Reference 39. [k] Reference 22. [l] Reference 27. [m] Reference 40. [n] Reference 31(b). [m] Δ represents the difference between frequencies calculated with the standard 6-31G(d,p) and the BC6-31G(d,p) basis sets.

Table 4. Mean Absolute Deviations for Frequencies (in cm^{-1})[a]

comparison (unscaled frequencies)	MAD	comparison (scaled frequencies)	MAD
experiment vs		experiment vs	
MP2/6-31G(d,p)	94.2	MP2/6-31G(d,p)	24.4
MP2/BC6-31G(d,p)	105.4	MP2/BC6-31G(d,p)	35.4
B3LYP/6-31G(d,p)	40.1	B3LYP/6-31G(d,p)	29.6
B3LYP/BC6-31G(d,p)	57.8	B3LYP/BC6-31G(d,p)	29.3
MP2/6-31G(d,p) vs		MP2/6-31G(d,p) vs	
MP2BC/6-31G(d,p)	16.4	MP2BC/6-31G(d,p)	15.3
B3LYP/6-31G(d,p)	58.9	B3LYP/6-31G(d,p)	22.7
B3LYP/BC6-31G(d,p) *vs*		B3LYP/BC6-31G(d,p) vs	
B3LYP/6-31G(d,p)	20.3	B3LYP/6-31G(d,p)	19.5
MP2/BC6-31G(d,p)	51.8	MP2/BC6-31G(d,p)	18.9

[a] A total of 73 frequencies were used to calculate the MAD between experimental and calculated frequencies, and 145 frequencies were used to calculate the MAD between the calculated frequencies.

It is interesting to note that after scaling the MAD are now similar for all levels of theory and basis sets.

The frequency scaling factors for first and second row elements are 0.9608 and 0.9370 at B3LYP/6-31G(d,p) and MP2/6-31G(d,p), respectively.[12,13] Using the 73 experimental frequencies available for compounds containing third row elements scaling factors were calculated by dividing the experimental frequencies with the corresponding calculated frequencies and then taking their average. The scale factors were found to be 0.9408 and 0.9246 at B3LYP/6-31G(d,p) and B3LYP/BC6-31G(d,p), respectively, and 0.8982 and 0.8926 at MP2/6-31G(d,p) and MP2/BC6-31G(d,p), respectively. These scaling factors indicate that in general frequencies calculated for compounds involving third row elements tend to be generally higher than those calculated for compounds containing first and second row elements.

3.3. Thermodynamic Properties for the Isogyric Reactions Involving Third Row Elements. The thermodynamic properties for reactions 1 and 2 are listed in Table 5.

For all X, X = Ga, Ge, As, Se, and Br, reaction 1 is exothermic with G3MP2 enthalpies of −2.3, −9.2, −29.6, −46.7, and −56.0 kJ mol^{-1}, respectively. From Figure 1, it is interesting to note that all levels predict that the enthalpy of reaction becomes more exothermic in going from Ga to Br. The G3MP2 free energies of reaction for X = Ge, As, Se and Br are exergonic with values of −9.2, −29.9, −44.5, and −54.3 kJ mol^{-1}, while for X = Ga the reaction is slightly endergonic with a G3MP2 free energy of 5.0 kJ mol^{-1}. For X = Ga, Ge, and As, reaction 2 is exothermic with G3MP2 enthalpies of −23.8, −25.3, and −14.2 kJ mol^{-1}, respectively, while for X = Se and Br, reaction 2 is endothermic with ΔH of 13.0 and 43.6 kJ mol^{-1}, respectively. From Figure 2, we see that in this case the enthalpy of reaction becomes more endothermic in going from Ga to Br for all

Table 5. Thermodynamic Properties for the Reactions 1 and 2 (in kJ mol^{-1}) at 298.15 K

	$CH_3GaH_2 + HCN \rightarrow CH_3CN + GaH_3$							$SiH_3GaH_2 + HCN \rightarrow SiH_3CN + GaH_3$						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]
HF	−22.0	−25.5	−18.6	−11.5	−14.7	−8.3	−10.8	−29.0	−36.4	−29.5	−24.4	−31.6	−24.6	−4.8
MP2(P)[b]	−4.1	−6.8	−0.9	0.2	−2.3	2.7	−4.5	−22.6	−28.4	−21.5	−22.4	−25.6	−28.4	−2.8
B3LYP(P)[b]	−14.5	−17.9	−11.8	−8.1	−8.7	−11.1	−9.2	−17.9	−24.3	−18.9	−16.1	−22.3	−17.5	−2.0
G3MP2	−0.8	−2.3	5.0					−22.3	−23.8	−19.3				

	$CH_3GeH_3 + HCN \rightarrow CH_3CN + GeH_4$							$SiH_3GeH_3 + HCN \rightarrow SiH_3CN + GeH_4$						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]
HF	−24.3	−27.4	−27.8	−5.3	−8.7	−9.3	−18.7	−31.1	−36.5	−36.4	−14.7	−19.7	−19.6	−16.8
MP2(P)[b]	−11.1	−13.4	−14.1	1.4	−0.4	−1.3	−13.0	−24.6	−28.5	−28.6	−16.7	−19.7	−19.7	−8.8
B3LYP(P)[b]	−22.2	−25.6	−25.6	−10.2	−13.0	−13.5	−12.6	−21.6	−26.4	−26.0	−14.1	−18.0	−17.7	−8.4
G3MP2	−8.6	−9.2	−9.2					−24.4	−25.3	−27.5				

	$CH_3AsH_2 + HCN \rightarrow CH_3CN + AsH_3$							$SiH_3AsH_2 + HCN \rightarrow SiH_3CN + AsH_3$						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]
HF	−41.2	−45.4	−46.0	−25.5	−30.3	−31.3	−15.1	−19.4	−24.9	−25.6	12.0	5.2	4.3	−30.1
HF(P)[b]	−40.6	−44.6	−45.2	−30.8	−34.7	−35.7	−9.9	−37.7	−24.8	−25.4	3.7	−1.8	−2.4	−23.0
MP2	−29.8	−33.2	−34.0	−15.9	−19.9	−21.1	−13.3	−9.7	−14.0	−14.9	19.2	13.7	12.5	−27.7
MP2(P)[b]	−29.2	−32.3	−33.0	−20.0	−23.0	−24.2	−9.3	−11.8	−15.8	−16.7	10.0	5.8	4.9	−21.6
B3LYP	−41.8	−45.7	−46.1	−27.0	−31.4	−32.2	−14.3	−14.3	−19.2	−19.3	14.1	7.9	7.3	−27.1
B3LYP(P)[b]	−39.1	−43.2	−43.6	−30.0	−33.8	−34.6	−9.4	−12.3	−17.3	−17.3	8.3	3.1	2.7	−20.4
G3MP2	−29.6	−29.6	−29.9					−14.1	−14.2	−17.3				

	$CH_3SeH + HCN \rightarrow CH_3CN + SeH_2$							$SiH_3SeH + HCN \rightarrow SiH_3CN + SeH_2$						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]
HF	−50.9	−56.6	−54.5	−41.5	−47.3	−45.4	−9.3	15.9	8.8	10.1	46.9	39.2	40.3	−30.4
HF(P)[b]	−53.5	−59.0	−56.9	−49.1	−54.2	−52.3	−4.8	12.9	6.0	7.3	37.1	30.4	31.6	−24.4
MP2	−39.8	−44.5	−42.7	−33.1	−37.7	−36.0	−6.8	22.7	16.8	17.7	50.7	44.2	45.0	−27.4
MP2(P)[b]	−42.9	−47.3	−45.4	−39.9	−43.9	−42.2	−3.4	17.8	12.2	13.1	40.4	34.9	35.9	−22.7
B3LYP	−49.4	−54.6	−52.4	−40.6	−45.5	−43.3	−9.1	15.6	9.1	10.5	43.4	36.7	38.2	−27.6
B3LYP(P)[b]	−49.6	−55.0	−52.7	−45.6	−50.5	−47.9	−4.5	14.9	8.3	9.8	36.3	30.2	32.7	−21.9
G3MP2	−47.1	−46.7	−44.5					12.3	13.0	11.7				

	$CH_3Br + HCN \rightarrow CH_3CN + HBr$							$SiH_3Br + HCN \rightarrow SiH_3CN + HBr$[e]						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]	Δ***E***	Δ***H***	Δ***G***	Δ***E***	Δ***H***	Δ***G***	Δ(Δ***H***)[a]
HF	−49.3	−55.4	−53.6	−45.2	−51.4	−49.6	−4.0	58.2	50.1	50.6	82.0	73.5	73.9	−23.4
HF(P)[b]	−56.2	−62.0	−60.1	−55.6	−61.2	−59.3	−0.8	51.0	43.3	43.9	71.0	63.2	63.6	−19.9
MP2	−42.4	−47.3	−45.8	−41.0	−45.8	−44.3	−1.5	58.4	51.8	51.9	79.6	72.6	72.6	−20.8
MP2(P)[b]	−50.5	−55.1	−53.5	−51.0	−55.4	−53.8	0.3	49.5	43.1	43.3	68.7	62.1	62.2	−19.0
B3LYP	−48.8	−54.1	−52.2	−44.5	−49.9	−48.0	−4.2	50.7	43.6	44.1	72.1	64.6	64.9	−21.0
B3LYP(P)[b]	−52.9	−58.1	−56.2	−52.2	−57.2	−55.3	−0.9	46.2	39.3	39.8	63.8	56.8	57.2	−17.5
G3MP2	−56.9	−56.0	−54.3	−57.1	−56.1	−54.5	0.1	42.0	43.6	43.8	40.4	42.0	42.1	1.6
exptl		−63.2[c]							29.6[d]					

[a] Δ(Δ*H*) represents the difference between enthalpies of reaction calculated with the standard 6-31G and the BC6-31G basis sets. [b] Represents 6-31G(d,p) basis set. [c] The value is calculated from experimental ΔH_f of CH_3Br, HCN, CH_3CN, and HBr given in Table 9. [d] The value is calculated from experimental ΔH_f of SiH_3Br, HCN, SiH_3CN, and HBr given in Table 9. [e] The thermodynamic properties are taken from ref 7.

levels of theory. Similarly, the free energies are exergonic for X = Ga, Ge, and As, with values of −19.3, −27.5, and −17.3 kJ mol^{-1}, respectively, and endergonic for X = Se and Br, with values of 11.7 and 43.8 kJ mol^{-1} at G3MP2.

For the reactions with CH_3Br and SiH_3Br, the G3MP2 enthalpies and free energies are calculated using both the standard 6-31G(d) and BC6-31G(d) basis sets. The G3MP2 energies calculated using the standard 6-31G(d)

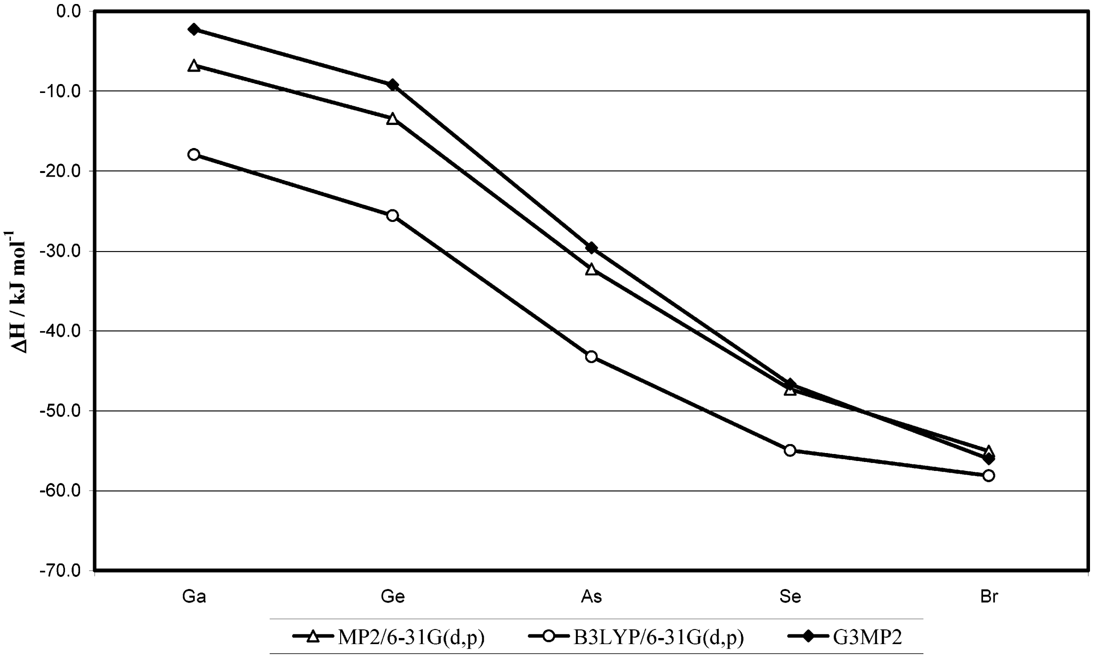


Figure 1. Enthalpies of reaction 1 calculated at different levels of theory with the standard 6-31G(d,p) basis set.

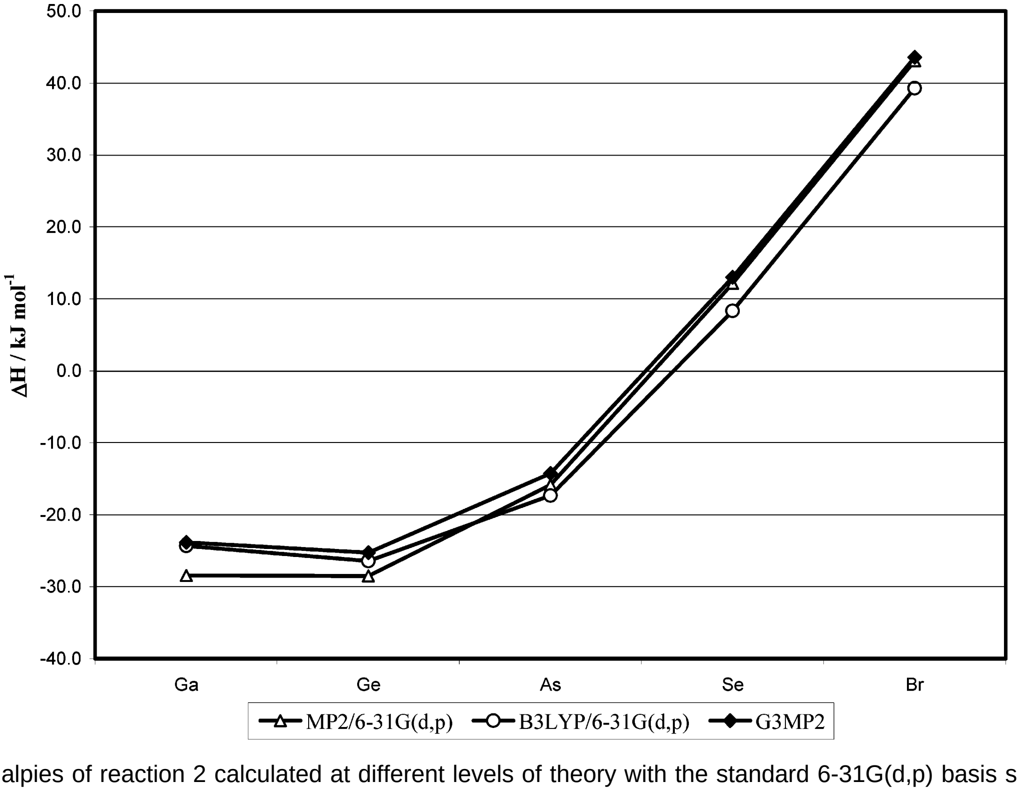


Figure 2. Enthalpies of reaction 2 calculated at different levels of theory with the standard 6-31G(d,p) basis set.

and the BC6-31G(d) basis set differ by only 0.2 kJ mol^{-1} for the reaction with CH_3Br and 1.6 kJ mol^{-1} with SiH_3Br (Table 5). Experimental enthalpies of reaction estimated from the heats of formation of the individual species (Table 9, to be discussed) are only available for the reaction with CH_3-Br and SiH_3Br. The G3MP2 enthalpies for both these two reactions agree reasonably well with experiment deviating by 7 kJ mol^{-1} and 14 kJ mol^{-1}, respectively. Although in some reactions addition of p-polarization functions to hydrogen gives better thermodynamic values, overall polarization functions have little effect on the thermodynamics.

Figures 3 (reaction 1) and 4 (reaction 2) represent the differences between the G3MP2 enthalpies from the enthalpies calculated at the MP2 and B3LYP levels of theory using both the standard 6-31G(d,p) and BC6-31G(d,p) basis sets. From Figure 3, it is clear that when X = Br, the error in the enthalpies calculated at the MP2 and B3LYP levels of theory is small for both the basis sets. This is similar to our previous investigation on the bromination of alkenes.[8] However, the errors in enthalpies calculated at B3LYP/6-31G(d,p) are slightly larger for X = Ga, Ge, As, and Se. For all X in reaction 2, the enthalpies of reaction calculated at both the

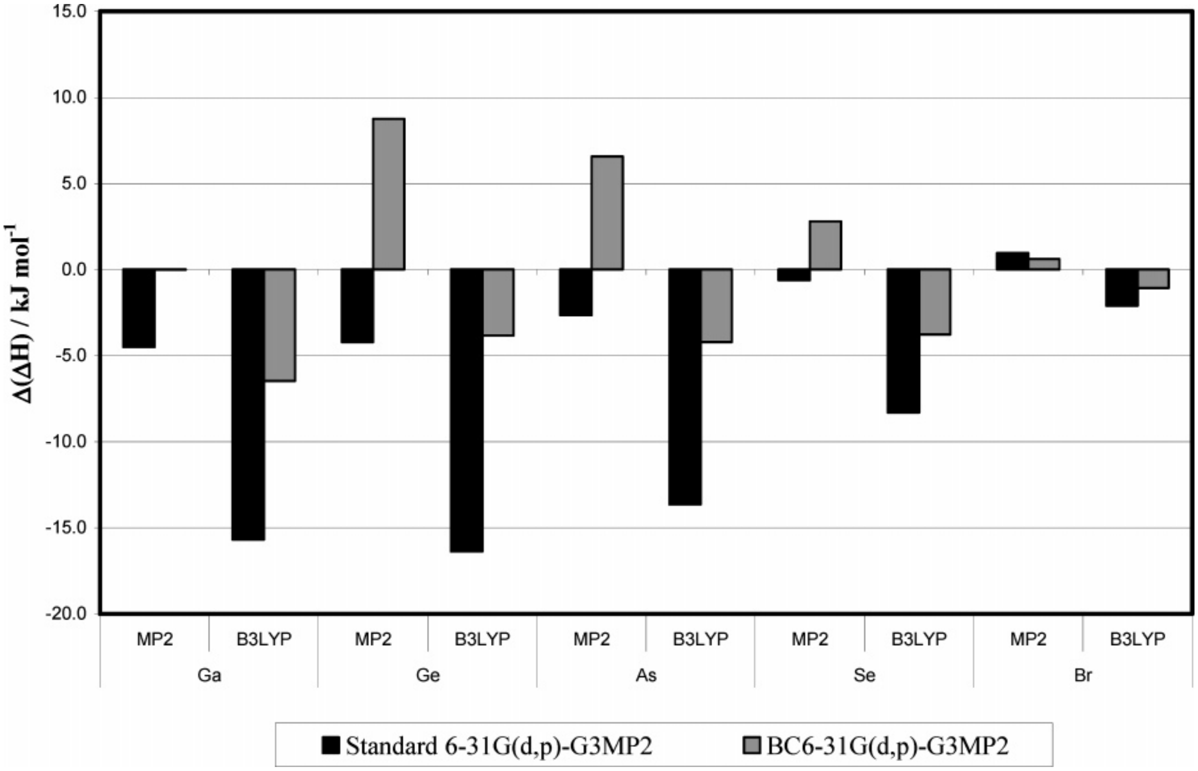


Figure 3. Difference between enthalpies of reaction 1 calculated at the MP2 and B3LYP levels of theory with G3MP2.

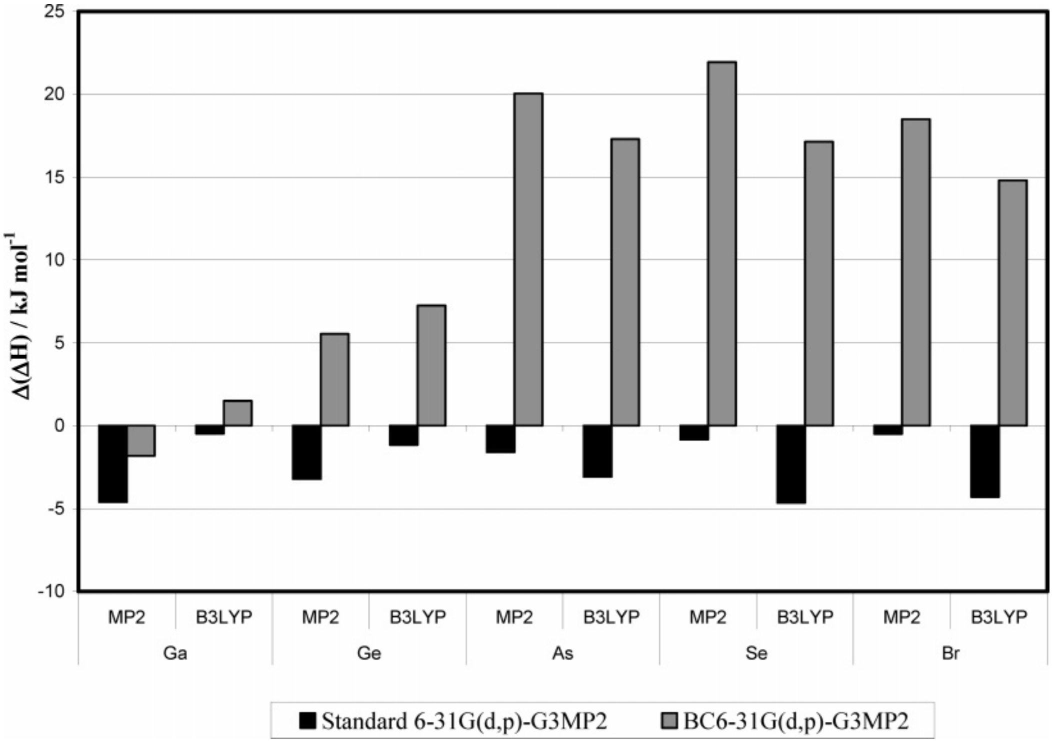


Figure 4. Difference between enthalpies of reaction 2 calculated at the MP2 and B3LYP levels of theory with G3MP2.

MP2 and B3LYP levels of theory using the standard 6-31G-(d,p) basis set are in excellent agreement with the G3MP2 enthalpies (all within 5 kJ mol^{-1}), while the BC6-31G(d) basis set performs especially poorly for X = Ge, As, Se, and Br. It is important to mention here that for the reaction of HCN with SiH_3AsH_2 the enthalpy calculated by the BC6-31G basis set is found to be endothermic, while with standard 6-31G, it is found to be exothermic in agreement with the G3MP2 level of theory (Table 5).

Both reactions 1 and 2 involved HCN as one of the reactants. To see the effect of second row elements on reaction thermodynamics, two more reactions, reaction 3 ($CH_3Br + HCl \rightarrow CH_3Cl + HBr$) and reaction 4 ($SiH_3Br + HCl \rightarrow SiH_3Cl + HBr$) are considered. The thermodynamic

Table 6. Thermodynamic Properties for the Reactions 3 and 4 (in kJ mol^{-1}) at 298.15 K

	$CH_3Br + HCl \rightarrow CH_3Cl + HBr$							$SiH_3Br + HCl \rightarrow SiH_3Cl + HBr$						
	6-31G(d)			BC6-31G(d)				6-31G(d)			BC6-31G(d)			
level	Δ*E*	Δ*H*	Δ*G*	Δ*E*	Δ*H*	Δ*G*	Δ(Δ*H*)[a]	Δ*E*	Δ*H*	Δ*G*	Δ*E*	Δ*H*	Δ*G*	Δ(Δ*H*)[a]
HF	1.0	0.5	−2.4	5.1	4.5	1.7	−4.0	−5.4	−6.7	−6.8	18.4	16.7	16.5	−23.4
HF(P)[b]	6.0	5.4	2.6	6.6	6.2	3.4	−0.8	−1.2	−2.4	−2.4	18.7	17.5	17.4	−19.9
MP2	2.5	2.1	−0.7	3.9	3.6	0.8	−1.5	−1.7	−2.9	−2.9	19.5	17.9	17.8	−20.8
MP2(P)[b]	8.9	8.4	5.6	8.4	8.1	5.3	0.3	4.0	2.9	2.9	23.2	21.8	21.8	−18.9
B3LYP	2.3	1.9	−0.9	6.6	6.2	3.3	−4.3	−5.2	−6.1	−6.1	16.2	14.9	14.7	−21.0
B3LYP(P)[b]	7.0	6.5	3.7	7.7	7.4	4.7	−0.9	−1.3	−2.3	−2.3	16.3	15.2	15.1	−17.5
G3MP2	11.3	11.1	8.3	11.2	11.0	8.2	0.1	−2.4	−2.7	−2.7	−4.0	−4.3	−1.7	−1.6
exptl		6.5[c]							−7.7[d]					

[a] Δ(Δ*H*) represents the difference between enthalpies of reaction calculated with the standard 6-31G and the BC6-31G basis sets. [b] Represents 6-31G(d,p) basis set. [c] The value is calculated from the experimental ΔH_f of CH_3Br, HCl, CH_3Cl, and HBr given in Table 9. [d] The value is calculated from the experimental ΔH_f of SiH_3Br, HCl, SiH_3Cl, and HBr given in Table 9.

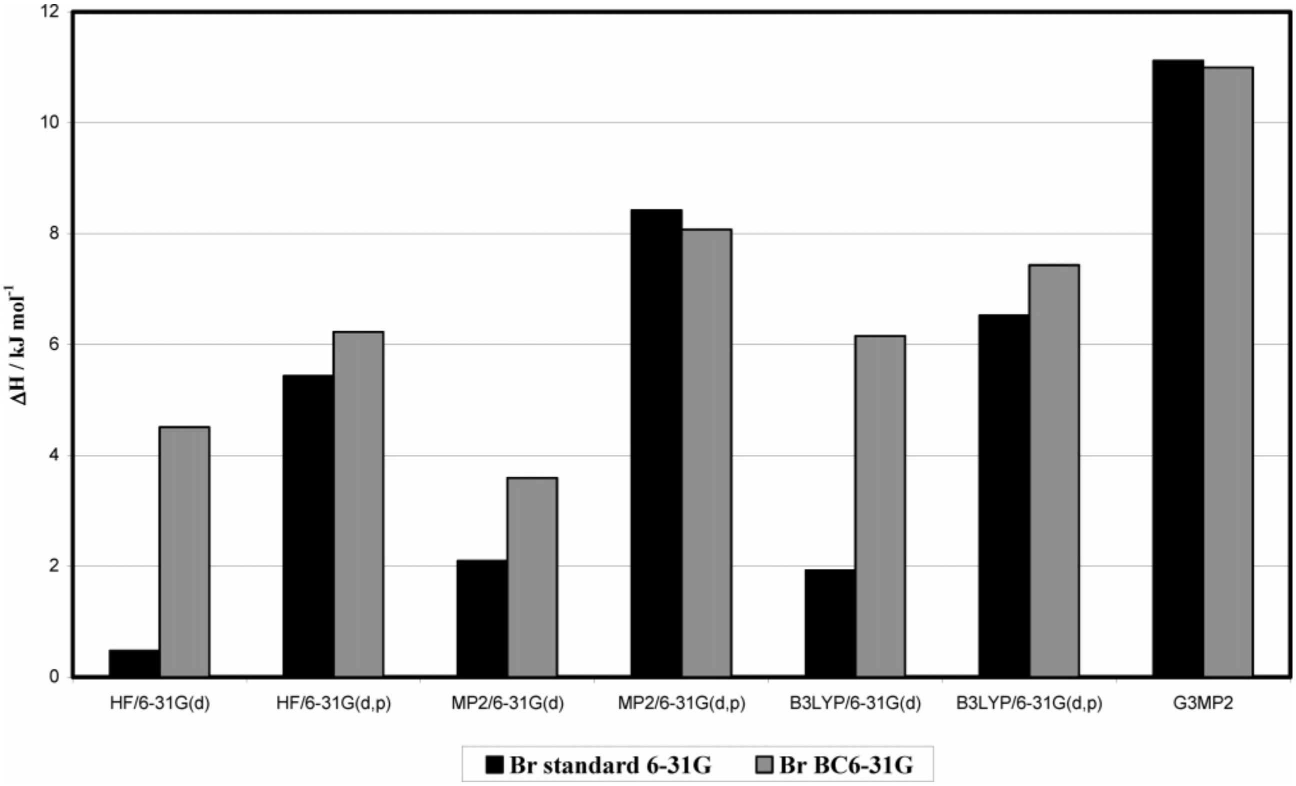


Figure 5. Enthalpy of reaction for $CH_3Br + HCl \rightarrow CH_3Cl + HBr$ calculated at different levels of theory and basis sets.

properties for reactions 3 and 4 are listed in Table 6, and the plots of reaction enthalpies vs theory/basis set are given in Figure 5 for reaction 3 and Figure 6 for reaction 4. G3MP2 enthalpies calculated with the standard 6-31G(d) basis set are in excellent agreement with the G3MP2 enthalpies calculated with the BC6-31G(d) basis set, differing by only 0.1 kJ mol^{-1} for reaction 3 and 1.6 kJ mol^{-1} for reaction 4. The G3MP2 enthalpies were also found to agree well with experiment differing by no more than 5 kJ mol^{-1}. The G3MP2 enthalpies calculated with both the standard 6-31G(d) and BC6-31G(d) basis set are found to be endothermic for reaction 3 and exothermic for reaction 4. For reaction 3, the HF, MP2, and B3LYP enthalpies calculated using the standard 6-31G and the BC6-31G basis set are in excellent agreement, differing by no more than 4.3 kJ mol^{-1}. In this case, all enthalpies of reaction are in good agreement with both the G3MP2 and experimental values. However, for reaction 4, the differences between the enthalpies of reaction calculated with the standard 6-31G and the BC6-31G basis set are large, ranging from 17.5 to 23.4 kJ mol^{-1}. The reaction enthalpies calculated with the standard 6-31G basis set are found to be exothermic (except for MP2/6-31G(d,p)), while the reaction enthalpies obtained by BC6-31G are endothermic for all levels of theory and basis sets investigated (Table 6 and Figure 6). In this case, the enthalpies calculated with the BC6-31G basis set are in poor agreement with both the G3MP2 and the experimental values. Therefore, the choice of basis set is extremely important for reactions involving both second and third row elements.

For reactions 2 and 4, both involving Si, the standard 6-31G basis set predicts better reaction enthalpies and free energies than the BC6-31G basis set. It would be interesting to see if the same result is found for other second row elements. Therefore, thermodynamic properties for reaction

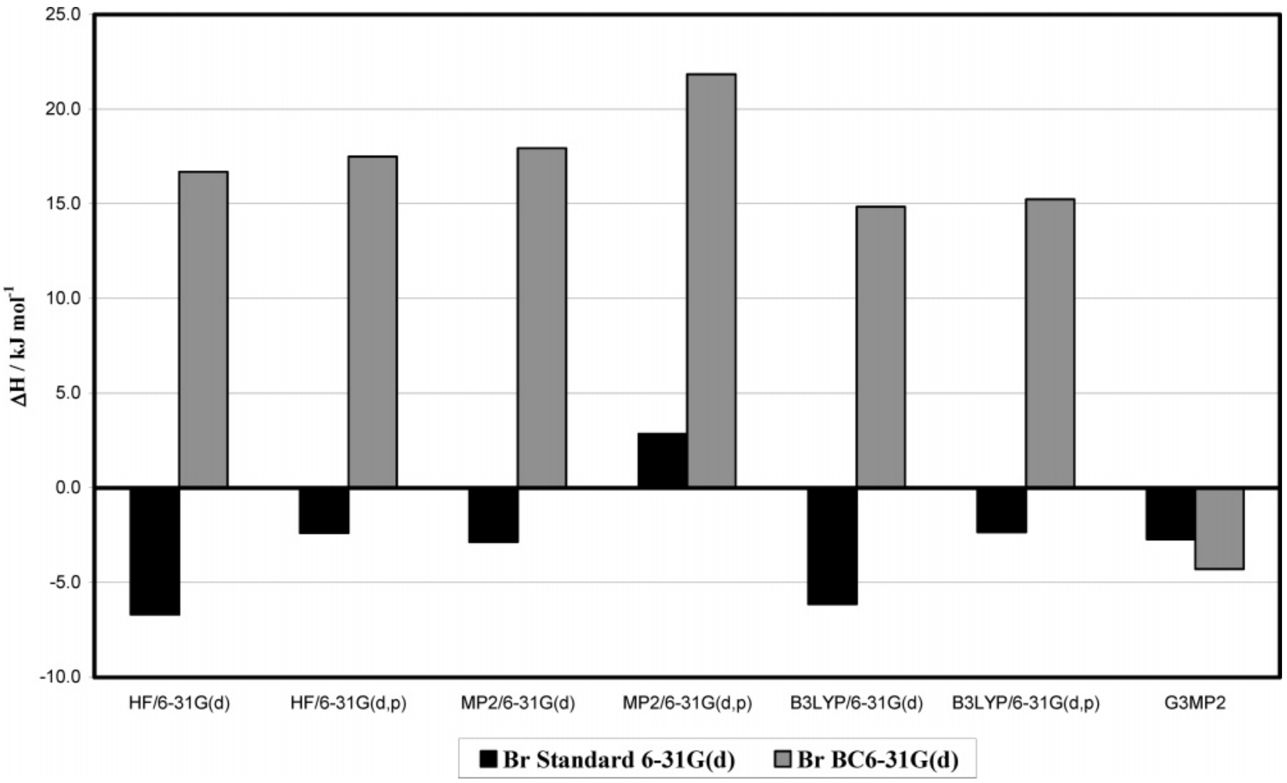


Figure 6. Enthalpy of reaction for $SiH_3Br + HCl \rightarrow SiH_3Cl + HBr$ calculated at different levels of theory and basis sets.

Table 7. Thermodynamic Properties for the Reaction 5 (in kJ mol^{-1}) at 298.15 K

	6-31G(d)			BC6-31G(d)			
level	Δ*E*	Δ*H*	Δ*G*	Δ*E*	Δ*H*	Δ*G*	Δ(Δ*H*)[a]
HF	18.8	11.4	12.8	34.4	26.8	28.0	−15.4
HF(P)[b]	11.9	4.9	6.3	23.7	16.7	18.0	−11.8
MP2	15.0	9.0	10.0	29.4	23.1	24.0	−14.1
MP2(P)[b]	6.3	0.6	1.6	18.6	12.8	13.8	−12.2
B3LYP	17.1	10.7	12.0	30.2	23.5	24.7	−12.8
B3LYP(P)[b]	12.9	6.5	7.9	22.1	15.8	17.1	−9.3
G3MP2	4.5	5.6	6.8				

[a] Δ(Δ*H*) represents the difference between enthalpies of reaction calculated with the standard 6-31G and the BC6-31G basis sets. [b] Represents the 6-31G(d,p) basis set.

5, $PH_2Br + HCN \rightarrow PH_2CN + HBr$, are calculated using both the standard 6-31G and the BC6-31G basis sets, and the values are given in Table 7. The plot of reaction enthalpies vs level of theory/basis set is shown in Figure 7. Differences in enthalpies calculated with the standard 6-31G and the BC6-31G basis sets range from 9.3 to 15.4 kJ mol^{-1} depending on the level of theory. Like reaction 4, SiH_3Br, the reaction enthalpies and free energies calculated with the standard 6-31G basis set is in better agreement with G3MP2 values (Table 7 and Figure 7).

Mean Absolute Deviations (MAD) of the Reaction Enthalpies. The mean absolute deviations for the reaction enthalpies involving first and third row elements, reaction 1, and for reactions involving first, second, and third row elements, reactions 2 and 5, are calculated at different levels of theory and basis sets from G3MP2 enthalpies, and the values are given in Table 8. The MAD for enthalpies of reactions involving first and third row elements are not significantly affected by the change of basis set, ranging from 2.6 to 13.5 kJ mol^{-1} for the standard 6-31G basis set and 1.8 to 5.8 kJ mol^{-1} for the BC6-31G basis set. The MAD are slightly higher at B3LYP/6-31G(d,p) and HF/6-31G(d). On the other hand, the MAD for the reaction enthalpies involving first, second, and third row elements (reactions 2 and 5) are significantly larger for the BC6-31G basis set at all levels of theory investigated, ranging from 10.1 to 18.4 kJ mol^{-1}. The MAD for the standard 6-31G basis set range from only 2.3 to 8.5 kJ mol^{-1} depending on the level of theory. Therefore, although the Binning-Curtiss and standard basis sets perform almost identically for reactions involving only first and third row elements, the standard basis set performs much better for reactions involving first, second, and third row elements. These results indicate that the BC6-31G basis set for third row elements is improperly balanced relative to the standard 6-31G basis set used for first and second row elements. The imbalance would result in basis set superposition error and basis set incompleteness error. The extra basis functions (3d) for the standard basis set are evidently playing a significant role, especially when bonding between second and third row elements is present.

3.4. Exploring Heats of Formation (Δ*H*$_f$). No experimental or theoretical heats of formation (Δ*H*$_f$) have been reported for CH_3SeH, SiH_3SeH, CH_3AsH_2, SiH_3AsH_2, CH_3GeH_3, and SiH_3GeH_3. In this study, the enthalpies for reactions 1 and 2 for all X, X = Ga, Ge, As, Se, and Br, have been obtained. The Δ*H*$_f$ values obtained in this study are given in Table 9. From the G3MP2 enthalpies of reaction and the most recent and reliable experimental heats of formation for CH_3CN, SiH_3CN, SeH_2, AsH_3, GeH_4, HCN, Δ*H*$_f$ for CH_3SeH, SiH_3SeH, CH_3AsH_2, SiH_3AsH_2, CH_3GeH_3, and SiH_3GeH_3 are calculated to be 18.3, 18.0, 38.4, 82.4,

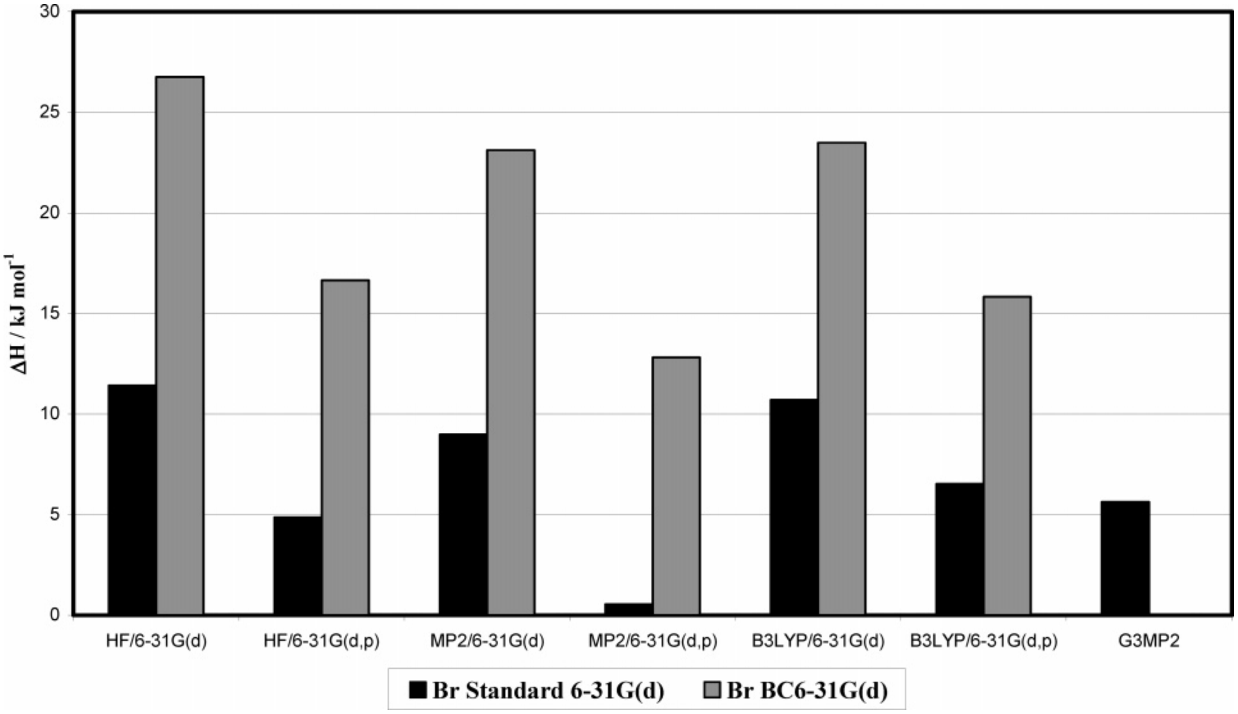


Figure 7. Enthalpy of reaction for PH_2Br + HCN → PH_2CN + HBr calculated at different levels of theory and basis sets.

Table 8. Mean Absolute Deviations for the Enthalpies of Reaction Involving First and Third Row Elements, Reaction 1, and First, Second, and Third Row Elements, Reactions 2 and 5 (in kJ mol^{1})

	reaction 1		reactions 2 and 5	
theory	6-31G(d)	BC6-31G(d)	6-31G(d)	BC6-31G(d)
HF	13.5	3.8	8.5	18.4
HF(P)[b]	6.7	3.6	3.1	10.1
MP2	2.9	5.8	2.6	17.6
MP2(P)[b]	2.6	3.8	2.6	12.5
B3LYP	5.2	1.8	2.3	14.1
B3LYP(P)[b]	11.2	3.9	2.4	11.1

[a] MAD is calculated from G3MP2 enthalpies. [b] Represents the 6-31G(d,p) basis set.

Table 9. Heats of Formation (ΔH_f) (in kJ mol^{-1}) at 298.15 K[a]

molecules	experiment	present work	molecules	present work[a]
CH_3Br	−38.0 ± 1.3[b]	−37.8,[n] −38.7[p]	CH_3SeH	18.3
HCN	131.67[c]	131.8,[n] 131.9[o]	SiH_3SeH	18.0
CH_3CN	74.04 ± 0.37[d]	73.9[n]	CH_3AsH_2	38.4
HBr	−36.2[e,f]	−36.4,[n] −36.5,[o] −35.5,[p] −31.4[q]	SiH_3AsH_2	82.4
SiH_3Br	−78.24[g]	−78.0,[o] −83.0[q]	CH_3GeH_3	41.9
SiH_3CN	133.5[h]	130.1[o]	SiH_3GeH_3	117.4
SeH_2	29.2 ± 0.8[i]			
AsH_3	66.4 ± 1[j,k]			
GeH_4	90.3 ± 2[l,m]			
CH_3Cl	−83.68[g]	−83.01[p]		
HCl	−92.31[g]	−93.0,[p] −97.1[q]		
SiH_3Cl	−141.84[g]	−137.1[q]		

[a] See text for explanation. [b] Reference 41, [c] Reference 42. [d] Reference 43. [e] Reference 44. [f] Reference 45. [g] Reference 46. [h] Reference 7 (obtained from experimental heats of formation and calculated heat of reaction). [i] Reference 47. [j] Reference 48. [k] Reference 49. [l] Reference 50. [m] Reference 51. [n] Calculated using the enthalpy of reaction for CH_3Br + HCN → CH_3CN + HBr and experimental ΔH_f values for CH_3Br, HCN, CH_3CN, and HBr. [o] Calculated using the enthalpy of reaction for SiH_3Br + HCN → SiH_3CN + HBr and experimental ΔH_f values for SiH_3Br, HCN, HBr, and SiH_3CN. [p] Calculated using the enthalpy of reaction for CH_3Br + HCl → CH_3Cl + HBr and experimental ΔH_f values for CH_3Br, HCl, CH_3Cl, and HBr. [q] Calculated using the enthalpy of reaction for SiH_3Br + HCl → SiH_3Cl + HBr and experimental ΔH_f values for SiH_3Br, HCl, SiH_3Cl, and HBr.

41.9, and 117.4 kJ mol^{-1}, respectively. Heats of formation were also calculated for HCN, CH_3CN, SiH_3CN, HBr, CH_3-Br, SiH_3Br, CH_3Cl, HCl and SiH_3Cl, for which reliable ΔH_f values are available for comparison. The ΔH_f for CH_3Br, HCN, CH_3CN, and HBr are calculated using the G3MP2 enthalpy of reaction for CH_3Br + HCN → CH_3CN + HBr (ΔH=−56.0 kJ mol^{-1} at G3MP2) and the most recent experimental heats of formation for CH_3Br, HCN, CH_3CN, and HBr (given in Table 9). The resulting ΔH_f values are −37.8, 131.8, 73.9, and −36.4 kJ mol^{-1}, respectively, all values being in excellent agreement with experiment. Similarly, heats of formation for HCN, SiH_3Br, SiH_3CN, and HBr are calculated using the enthalpy of reaction for SiH_3Br + HCN → SiH_3CN + HBr (43.6 kJ mol^{-1} at G3MP2), along with experimental heats of formation for HCN, SiH_3Br, HBr, and SiH_3CN. The ΔH_f values are again in excellent agreement with experiment. Heats of formation of CH_3Br, HBr, SiH_3Br, CH_3Cl, HCl, and SiH_3Cl are also calculated using the enthalpy of reaction 3, CH_3Br + HCl → CH_3Cl + HBr (11.1 kJ mol^{-1} at G3MP2), and reaction 4, SiH_3Br + HCl → SiH_3Cl + HBr (−2.7 kJ mol^{-1} at G3MP2) and by using the experimental heats of formation of CH_3Br, HCl, CH_3Cl, HBr, SiH_3Br, and SiH_3Cl. The ΔH_f values obtained by reaction 3 is in excellent agreement with experiment, while the values obtained by using reaction 4 is also in reasonable agreement with experiment differing by no more than 4.8 kJ mol^{-1} from experiment. Therefore, these results provide further evidence that the G3MP2 enthalpies are very reliable for the reactions studied and proved to be useful in predicting the performance of the standard 6-31G and BC6-31G basis sets.

4. Conclusions

Computations were carried out in order to compare the standard and BC6-31G basis sets for thermodynamic properties, geometries, and frequencies. The performance of the standard 6-31G basis set compared to the BC6-31G basis set for a series of isogyric reactions containing third row elements, Ga, Ge, As, Se, and Br, was evaluated using G3MP2 theory. A comparison of the thermodynamic properties calculated with the standard 6-31G and the BC6-31G basis set with the G3MP2 energies revealed that for compounds with first row elements and third row elements, both basis sets perform equally well, while compounds with second and third row elements or with first, second, and third row, elements, the standard 6-31G basis set showed the best performance. Optimized geometries were also tabulated and compared for the standard 6-31G(d,p) and BC6-31G(d,p) basis sets. Geometric parameters calculated with both the basis sets were found to agree well with experiment, with errors similar to those found for compounds containing first and second row elements. Frequencies were also compared to experiment, and the unscaled B3LYP/6-31G(d,p) frequencies were found to be in better agreement with experiment (Table 4). MP2/6-31G(d,p) were also found to predict better frequencies than MP2/BC6-31G(d,p). Scaling the frequencies with standard scale factors lowers the MAD for all levels and basis sets studied suggesting that the standard scale factors for first and second row elements may also be used for third row elements. Calculations using the G3MP2 theory proved useful in determining the accuracy of the levels of theory and basis sets. When studying reactions involving heavy atoms, the choice of the basis set is crucial. As illustrated in this study, enthalpies of reaction can vary up to 30.4 kJ mol^{-1} at the B3LYP and MP2 levels of theory which in several cases may lead to predicting a reaction is endothermic when it is actually exothermic and vice versa. Since the standard 6-31G basis set performs very well with all the reactions, we recommend that the standard 6-31G basis set be used for calculations involving third row elements. It has also been shown that reaction enthalpies calculated at G3MP2, along with existing experimental data, can be used to calculate reliable heats of formation.

Acknowledgment. We are grateful to the Natural Sciences and Engineering Council of Canada (NSERC) for financial support and the Atlantic Computational Excellence Network (ACEnet) for computer time.

Supporting Information Available: Full geometries and energies of all structures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Curtiss, L. A.; Redfern, P. C. *J. Chem. Phys.* **2001**, *114*, 9287−9295.

(2) Binning, Jr., R. C.; Curtiss, L. A. *J. Comput. Chem.* **1990**, *11*, 1206.

(3) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision B.05*; Gaussian, Inc.: Wallingford, CT, 2004.

(4) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347−1363. GAMESS Version = Feb 22, 2006 (R5) from Iowa State University.

(5) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. *J. Comput. Chem.* **2001**, *22*, 976−984.

(6) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(7) Islam, S. M.; Hollett, J. W.; Poirier, R. A. *J. Phys. Chem. A* **2007**, *111*, 526−540.

(8) Islam, S. M.; Poirier, R. A. *J. Phys. Chem. A* In press.

(9) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703−4709.

(10) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650−7657.

(11) Curtiss, L. *Computational Thermochemistry on the Web*. http://chemistry.anl.gov/compmat/comptherm.htm (accessed Sept 24, 2007).

(12) National Institute of Standards and Technology. *Vibrational Frequency Scaling Factors on the Web.* http://srdata.nist.gov/cccbdb/vsf.asp (accessed Sept 24, 2007).

(13) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502−16513.

(14) Bartmess, J. E.; Hinde, R. J. *Can. J. Chem.* **2005**, *83*, 2005−2012.

(15) Inorganic (non carbon-containing) compounds. In *Tables of Interatomic Distances and Configuration in Molecules and Ions*, Spec. Publ. 18; Sutton, L. E., Ed.; The Chemical Society: London, U.K.; 1956−1959; pp M 1s−M 58s.

(16) Stevenson, D. P. *J. Chem. Phys.* **1940**, *8*, 285−287.

(17) Ohno, K.; Matsumura, H.; Endo, Y.; Hirota, E. *J. Mol. Spectrosc.* **1986**, *118*, 1−17.

(18) Stevenson, P. E.; Lipscomb, W. N. *J. Chem. Phys.* **1970**, *52*, 5343−5353.

(19) Breidung, J.; Thiel, W.; Demaison, J. *Chem. Phys. Lett.* **1997**, *266*, 515−520.

(20) Harjanto, H.; Harper, W. W.; Clouthier, D. J. *J. Chem. Phys*. **1996**, *105*, 10189−10200.

(21) Herzberg, G.; Verma, R. D. *Can. J. Phys*. **1964**, *42*, 395−432.

(22) Herzberg, G. *Molecular Spectra and Molecular Structure. III. Electronic Spectra and Electronic Structure of Polyatomic Molecules*; D. Van Nostrand: Princeton, NJ, 1967.

(23) Graner, G. *J. Mol. Spectrosc*. **1981**, *90*, 394−438.

(24) Duncan, J. L. *J. Mol. Struct.* **1970**, *6*, 447−456.

(25) Duncan, J. L.; Harvie, J. L.; McKean, D. C.; Cradock, S. *J. Mol. Struct*. **1986**, *145*, 225−242.

(26) Chadwick, D.; Millen, D. J. *J. Mol. Struct*. **1975**, *25*, 216−218.

(27) Harvey, A. B.; Wilson, M. K. *J. Chem. Phys.* **1966**, *45*, 678−688.

(28) Harvey, A. B.; Wilson, M. K. *J. Chem. Phys*. **1966**, *44*, 3535−3546.

(29) Mathews, S.; Duncan, J. L.; McKean, D. C.; Smart, B. A. *J. Mol. Struct.* **1997**, *413*, 553−573.

(30) Laurie, V. W. *J. Chem. Phys.* **1959**, *30*, 1210−1214.

(31) (a) Obenhammer, H.; Lobreyer, T.; Sundermeyer, W. *J. Mol. Struct*. **1994**, *323*, 125−128. (b) Jensen, J. O. *Spectrochim. Acta, Part A* **2003**, *59*, 3093−3102.

(32) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*; Van Nostrand Reinhold: New York, 1979.

(33) Shimanouchi, T. Molecular Vibrational Frequencies. In N*IST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J. Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD 20899, June 2005. http://webbook.nist.gov (accessed Oct 31, 2007).

(34) Straley, J. W.; Tindal, C. H.; Nielsen, H. H. *Phys. Rev.* **1942**, *62*, 161−165.

(35) Pullumbi, P.; Bouteiller, Y.; Manceron, L.; Mijoule, C. *Chem. Phys*. **1994**, *185*, 25−37.

(36) Wang, X.; Andrews, L. *J. Phys. Chem. A* **200**3, *107*, 11371−11379.

(37) Muller, J.; Sternkicker, H.; Bergmann, U.; Atakan, B. *J. Phys. Chem. A* **2000**, *104*, 3627−3634.

(38) Paplewski, P.; Beckers, H.; Burger, H. *J. Mol. Spectrosc*. **2002**, *213,* 69−78.

(39) Duncan, J. L.; Allan, A.; McKean, D. C. *Mol. Phys*. **1970**, *18*, 289−303.

(40) Lannon, J. A.; Weiss, G. S.; Nixon, E. R. *Spectrochim. Acta, Part A* **1970**, *26*, 221−233.

(41) Adams, G. P.; Carson, A. S.; Laye, P. G. *Trans. Faraday Soc*. **1966**, *62*, 1447−1449.

(42) Hansel, A.; Scheiring, C.; Glantschnig, M.; Lindinger, W.; Ferguson, E. E. *J. Chem. Phys*. **1998**, *109*, 1748−1750.

(43) An, X.; Mansson, M. *J. Chem. Thermodyn*. **1983**, *15*, 287−293.

(44) *CRC Handbook of Chemistry and Physics*; CRC: Boca Raton, FL, 1977−1978; Vol. 58, p D69.

(45) McBride, B. J.; Zehe, M. J.; Gordon, S. *NASA Glenn Coefficients for Calculating Thermodynamic Properties of Individual Species*; Glenn Research Center: Cleveland, OH, 2002; p 20.

(46) Chase, M. W. *NIST-JANAF Thermochemical Tables*, 4th ed.; Monograph 9. *J. Phys. Chem. Ref. Data* 1998; pp 1−1951.

(47) Gibson, S. T.; Greene, J. P.; Berkowitz, J. *J. Chem. Phys.* **1986**, *85*, 4815−4824.

(48) Berkowitz, J. *J. Chem. Phys*. **1988**, *89*, 7065−7076.

(49) Gunn, S. R.; Jolly, W. L.; Green, L. G. *J. Phys. Chem*. **1960**, *64*, 1334−1335.

(50) Ruscic, B.; Schwarz, M.; Berkowitz, J. *J. Chem. Phys.* **1990**, *92*, 1865−1875.

(51) Gunn, S. R.; Green, L. G. *J. Phys. Chem.* **1961**, *65*, 779−783.

CT700224J

JCTC Journal of Chemical Theory and Computation

Mechanism of Air Oxidation of the Fragrance Terpene Geraniol

Carina Bäcktorp,[†] Lina Hagvall,[‡] Anna Börje,[‡] Ann-Therese Karlberg,[‡]
Per-Ola Norrby,[*,§] and Gunnar Nyman[†]

Department of Chemistry, Physical Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden, Department of Chemistry, Dermatochemistry and Skin Allergy, Göteborg University, SE-412 96 Göteborg, Sweden, and Department of Chemistry, Organic Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden

Received June 20, 2007

Abstract: The fragrance terpene geraniol autoxidizes upon air exposure and forms a mixture of oxidation products, some of which are skin sensitizers. Reactions of geraniol with O_2 have been studied with DFT (B3LYP) and the computational results compared to experimentally observed product ratios. The oxidation is initiated by hydrogen abstraction, forming an allylic radical which combines with an O_2 molecule to yield an intermediate peroxyl radical. In the subsequent step, geraniol differs from previously studied cases, in which the radical chain reaction is propagated through intermolecular hydrogen abstraction. The hydroxy-substituted allylic peroxyl radical prefers an intramolecular rearrangement, producing observable aldehydes and the hydroperoxyl radical, which in turn can propagate the radical reaction. Secondary oxidation products like epoxides and formates were also considered, and plausible reaction pathways for formation are proposed.

Introduction

Contact allergy, caused by skin-penetrating compounds able to react with macromolecules in the skin to form antigens, is one of the most common health problems in the industrialized world. In Western Europe, an estimated 10−15% of the normal population suffers from contact allergies that upon prolonged or repeated contact with the offending agent result in allergic contact dermatitis. Fragrance compounds commonly cause contact allergies. Fragrances are ubiquitous in our environment, and not only cosmetics and toiletries contain fragrance materials but almost all household and occupational products are scented. The allergens are not always the fragrance compounds themselves, but rather degradation products formed upon prolonged storage in contact with air, for example, in scented products.

As part of a long-term project of identifying compounds that are not allergenic themselves but can form allergenic compounds upon air exposure, it has been shown how some common fragrance terpenes form allergenic hydroperoxides and secondary oxidation products upon exposure to air. The oxidation products were isolated and identified, and their allergenic effects were determined experimentally.[1−4]

In the mechanism for autoxidation of the unsaturated terpene linalool (**1**, Figure 1), oxidation was found to occur by abstraction of an allylic hydrogen, followed by combination with O_2 and radical chain propagation to yield allylic hydroperoxides as primary oxidation products.[5]

A recent study investigated the bimolecular reaction between an alkene and triplet oxygen, requiring a spin-state change to reach the singlet products.[6] However, in general, the formation of hydroperoxides via autoxidation is believed[5] to proceed through a radical chain process according to the following steps:

$$\text{Initiation: } RH \rightarrow R\cdot \qquad \text{(step 1)}$$

$$\text{Propagation: } R\cdot + O_2 \rightarrow ROO\cdot \qquad \text{(step 2a)}$$

$$ROO\cdot + R'H \rightarrow ROOH + R'\cdot \qquad \text{(step 2b)}$$

$$\text{Termination } 2R'\cdot \rightarrow \text{nonradical products} \qquad \text{(step 3)}$$

[*] Corresponding author e-mail: pon@chem.gu.se.
[†] Department of Chemistry, Physical Chemistry.
[‡] Department of Chemistry, Dermatochemistry and Skin Allergy.
[§] Department of Chemistry, Organic Chemistry.

10.1021/ct7001495 CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/03/2007

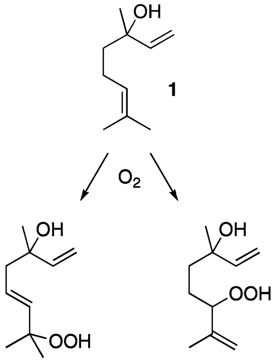


Figure 1. Linalool (**1**) and two hydroperoxides identified after air oxidation.

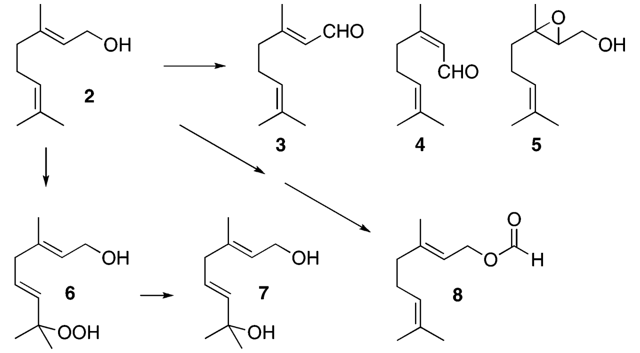


Figure 2. Geraniol (**2**) with identified air oxidation products.

Here 2R′• can be any combination of the radicals formed. We note that no step in the chain process requires a change of spin. The exothermic addition of oxygen to the radical, step 2a, is believed to occur without any barrier on the potential energy surface (i.e., it is diffusion-controlled, *vide infra*); hence, the rate and selectivity determining step of the propagation is the hydrogen atom abstraction, step 2b.

Geraniol (**2**, *trans*-3,7-dimethyl-2,6-octadiene-1-ol), an isomer of **1**, is an important fragrance terpene, widely used because of its fresh flowery odor. A recent investigation[7] of the air oxidation of geraniol (**2**, Figure 2) revealed that the reaction is substantially more complex than that of **1**, forming a mixture of products that include hydrogen peroxide, the aldehydes geranial (**3**) and neral (**4**), and epoxygeraniol (**5**), in addition to a hydroperoxide (**6**) related to those found in the linalool study,[4] and its secondary degradation product, the allylic alcohol **7**. Furthermore, the presence of geranyl formate (**8**) in the oxidation mixture must be rationalized by a postoxidation bimolecular transformation, since the additional carbon in the formate moiety has to come from another, degraded geraniol molecule. Studies of the skin-sensitizing potency according to the local lymph node assay in mice showed that air-exposed geraniol as well as several of the isolated oxidation products have a sensitizing potency significantly higher than that of pure geraniol, demonstrating the need for an increased understanding of the oxidation of fragrance terpenes.[7]

In an earlier investigation of geraniol, it could be concluded that the most easily abstracted allylic hydrogen is the one α to the hydroxyl, leading to the preferential formation of radical **A** (Figure 3).[7]

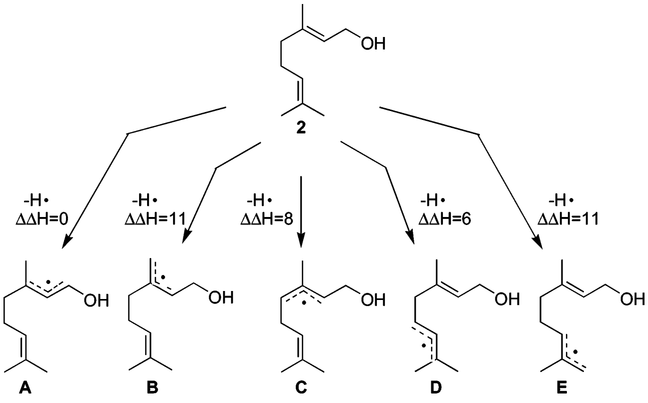


Figure 3. Illustration of five radicals that can be formed by hydrogen abstraction from geraniol. Enthalpy changes (in kcal mol^{-1}) for forming the various radicals are given relative to radical **A**.

Radical **D** is less stable than **A**, but some products derived from radical **D** were still observed (**6** and **7**, Figure 2). The three other radicals, **B**, **C**, and **E**, are even higher in energy, and indeed no oxidation products derived from any of these could be identified.[7] The formation of products from **D** is analogous to the previously investigated oxidation of linalool (**1**)[3−5] and will not be further discussed here. Instead, we will concentrate on the possible further reactions of radical **A**.

In this work, we report our theoretical investigation of the mechanism of oxidation of geraniol (**2**). In the current work, we focus upon the primary oxidation, which follows a radical chain process, forming the primary oxidation products in the presence of triplet oxygen. Secondary oxidation products are then formed in closed-shell processes for which there are ample precedents in the literature.

Computational Methods

All calculations were performed using unrestricted density functional theory (DFT) with the B3LYP functional[8] as implemented in Gaussian 03.[9] We utilized two different basis sets, optimizing all structures first with 6-31G(d,p), and then using the larger 6-311+G(2d,p) basis set. Harmonic vibrational frequencies were obtained for all structures (and both basis sets) in order to ensure the nature of the stationary points (saddle point or minimum), and also to estimate the thermodynamic contribution to the enthalpy and free energy at $T = 298$ K. Energies are reported as enthalpies at 0 K and at 298 K, and free energies at 298 K.

Results and Discussion

In the investigation of the formation and further reactions of radical **A** (Figure 3), we have chosen to use a smaller and less flexible molecule, 3-methyl-2-buten-1-ol (**9,** Figure 4), as the model for geraniol (**2**). This model includes all the features of **2** necessary for reproducing the stability of **A**, that is, the trisubstituted alkene and the allylic hydroxy functionality, but excludes the conformationally flexible isoprenoid moiety that is expected to stay constant in all investigated reactions. Depending on the conformation of the alcohol in the hydrogen abstraction step, two radicals

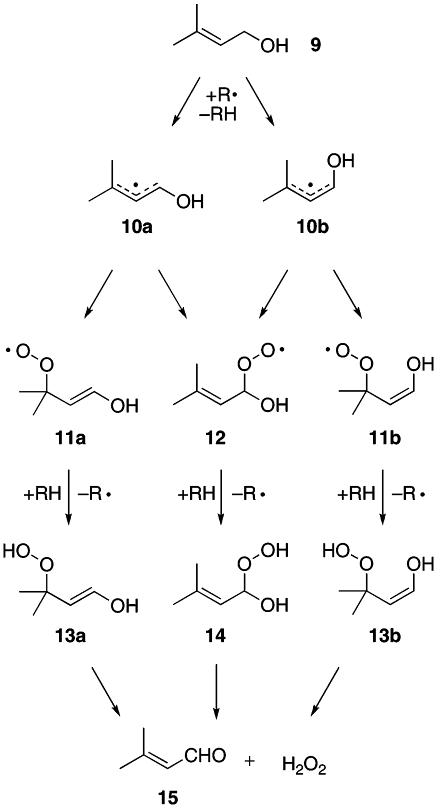


Figure 4. Propagation steps using the geraniol model **9**.

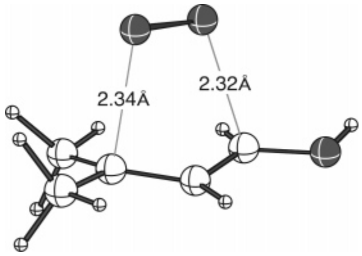


Figure 5. B3LYP/6-311+G(2d,p) TS for direct interconversion between **11a** and **12**.

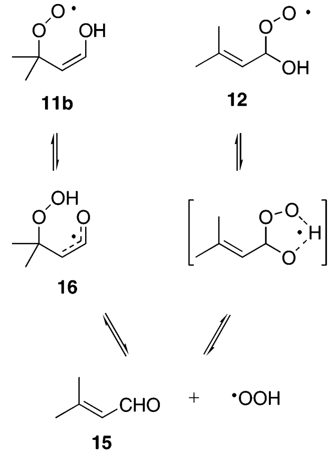


Figure 6. Alternative fragmentation paths for peroxyl radicals **11b** and **12**.

can be formed, **10a** and **10b** (Figure 4). For steric reasons, we expect the trans form, **10a**, to dominate, but both forms, and their potential interconversion pathways, have been included in the study.

Table 1. Calculated Standard Enthalpy Changes and Standard Gibbs Free Energy Changes in kcal mol^{-1} for the Reactions in Figure 4

	6-31G(d,p)			6-311+G(2d,p)		
	$\Delta H_0°$	$\Delta H_{298}°$	$\Delta G_{298}°$	$\Delta H_0°$	$\Delta H_{298}°$	$\Delta G_{298}°$
R· + O_2 → ROO· (step 2a)						
10a + O_2 → 11a	−16	−14	−2	−12	−13	0
10a + O_2 → 12	−21	−18	−7	−16	−17	−5
10b + O_2→ 11b	−23	−20	−7	−16	−17	−4
10b + O_2 → 12	−22	−20	−8	−17	−18	−7
ROO· + RH → ROOH + R· (step 2b)						
11a → 13a	−2	−3	−3	−4	−3	−4
12 → 14	−2	−3	−4	−5	−4	−6
11b → 13b	0	0	−1	−2	−1	−3
ROOH → aldehyde + H_2O_2						
13a → 15 + H_2O_2	5	3	−10	−1	0	−13
14 → 15 + H_2O_2	10	8	−4	4	5	−7
13b → 15 + H_2O_2	9	6	−8	0	1	−13

Table 2. Calculated Activation Energies, in kcal mol^{-1}

	6-31G(d,p)			6-311+G(2d,p)		
TS	$\Delta H_0°$	$\Delta H_{298}°$	$\Delta G_{298}°$	$\Delta H_0°$	$\Delta H_{298}°$	$\Delta G_{298}°$
11b → 16	6	5	6	6	6	7
12 → 15	3	3	3	4	4	4
11a → 12	10	10	10	9	9	9
11b → 12	17	17	15	15	15	16

No transition states (TSs) could be found for the first propagation step, the combination of radicals **10a** and **10b** with O_2. To verify that the addition is indeed barrier-less, we performed a geometry optimization starting with an O_2 molecule positioned perpendicular to the π-face of radical **10a**, with the closest oxygen−carbon distance set to 3 Å. In this optimization, the trust radius was strongly reduced, to 0.05 b, to ensure that the optimization sequence did not accidentally skip over a low barrier. Each step of the optimization was inspected, verifying that the steps were small and the energy decrease monotonous. The optimization proceeded as expected, and yielded structure **12**, showing that the addition can occur without an energy barrier.

The addition products **11** and **12** can potentially equilibrate by reverting to free O_2 and allylic radicals **10**. However, a direct [2,3] shift is also possible and was found to have an activation energy lower than that required for the dissociation of O_2. The transition state for direct conversion between **11a** and **12** is depicted in Figure 5. As can be seen, the TS is very symmetric, with forming and breaking C−O bonds of almost equal length. Interestingly enough, the TS structure is also very similar to one of the intermediate points in the slow optimization used to verify the barrier-less nature of the O_2 addition, indicating that it is also a potential branching point for the O_2 addition reaction.

In the last propagation step, peroxyl radicals **11**/**12** abstract a hydrogen from another species in solution, forming a new radical and the hydroperoxy species **13** and **14**. Each of these are expected to be in equilibrium with aldehyde **15** and hydrogen peroxide, which has also been detected in the

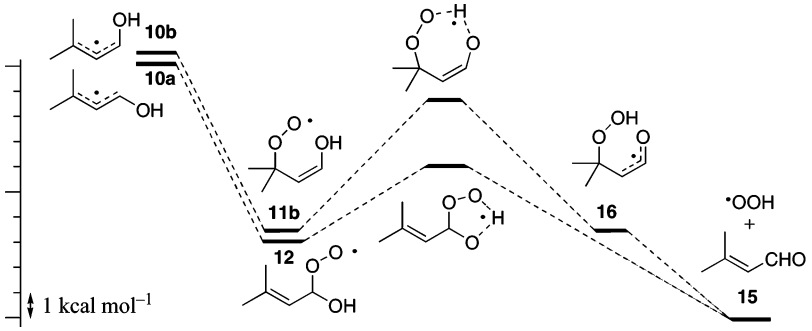


Figure 7. Free energy surface for the fragmentations depicted in Figure 6.

autoxidation sample. The calculated reaction energies and barriers for these steps are shown in Tables 1 and 2, respectively.

Analyzing first the effect of the two different basis sets, we can see that there are numerical differences, but that the qualitative picture is the same in both cases. Moreover, the change follows the expected trend. The basis superposition error (BSSE), which can be significant with smaller basis sets, will favor association, since a more compact arrangement of atoms allows the virtual orbitals from one fragment to fill out deficiencies in the orbital description of the neighboring fragment. Thus, with the larger basis set, the reaction between O_2 and allyl radicals becomes less exothermic, the proton-transfer steps are virtually unaffected, and the dissociation steps become more exothermic. We also want to point out that the BSSE to some extent compensates for an error in B3LYP, namely, the lack of proper treatment of dispersion forces. We are not aware of a full investigation of the relative magnitude of these effects, but in our experience with similar methods, a modest basis set frequently gives better agreement with experiments than the more extensive one. To conclude, we cannot be certain which of the two sets of data is in best agreement with experimental values, but it is reassuring that both sets give the same qualitative results. Since this is the case, we will perform additional calculations using the cheaper of the two methods, B3LYP/6-31G(d,p).

Looking at the two propagation steps, the initial combination of the allylic radical with O_2 (step 2a) occurs without a barrier on the potential energy surface (*vide supra*). For step 2b, the abstraction of a hydrogen atom from another molecule in solution forming hydroperoxides, we have investigated a model peroxyl radical, $CH_3OO\cdot$, reacting with the geraniol model **9** to form **10**, at the B3LYP/6-31G(d,p) level. For this step, we find an enthalpy of activation of 7 kcal mol^{-1}, and a free energy of activation of 18 kcal mol^{-1}. Thus, the reversion of step 2a, which is endergonic by 0−8 kcal mol^{-1},[10] is competitive with propagation. Peroxyl radical **11a** has no alternative forward reaction and may either revert to **10a** or isomerize to **12** to a significant extent. On the other hand, **11b** and **12** can follow alternative fragmentation paths due to the spatial proximity of the hydroxy group (Figure 6), in a heteroatom analogy to the known fragmentation of the ethylperoxyl radical.[11] As seen in the corresponding free energy surface, Figure 7, the intramolecular hydrogen transfer in **11b** to produce peroxy enolyl radical **16** is virtually isoergonic, with a moderate barrier. Intermediate **16** can then eliminate a hydroperoxyl radical in an exergonic process, to form aldehyde **15**. For the hydroxy-substituted peroxyl radical **12**, the intramolecular hydrogen transfer and elimination are concerted, forming the hydroperoxyl radical and free aldehyde with a very low barrier (Figure 8).

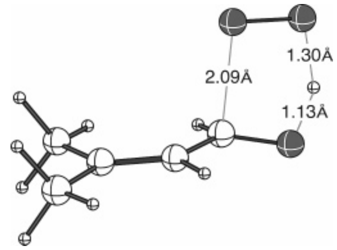


Figure 8. B3LYP/6-311+G(2d,p) TS for fragmentation of **12** to **15**.

Overall, fragmentation via **12** and intramolecular elimination is the preferred path, but as can be seen in Figure 7 and Table 2, when formed, peroxyl radical **11b** will prefer elimination via **16** over reversion to allylic radical **10** or isomerization to **12**.

In the full system starting from geraniol (**2**), we must also consider cis/trans isomerization, giving neral (**4**) in addition to geranial (**3**). In Figure 9, we have summarized the expected pathways and indicated intermediates where isomerization around the former double bond is feasible. For one peroxyl radical where no intramolecular hydrogen abstraction is possible (corresponding to **11a** in Figure 4), an equilibrium back to free O_2 and the allylic radical or a [2,3] shift is expected. Dotted arrows indicate intermolecular hydrogen abstraction followed by closed-shell fragmentation, as outlined in Figure 4. This process is expected to be disfavored compared with reversal and branching to a pathway allowing intramolecular hydrogen abstraction and fragmentation.

The hydroperoxyl radical produced by the fragmentations depicted in Figures 6 and 9 can participate in the radical chain propagation by abstracting a hydrogen atom from a molecule of geraniol (**2**). However, the reactive hydroperoxyl radical can also add to the double bond of geraniol, as shown for the model compound **9** in Figure 10. The addition is somewhat endergonic, but the subsequent ring closure to epoxy alcohol **17** is strongly exergonic. The liberated hydroxyl radical is highly reactive and will propagate the radical chain process by the abstraction of a hydrogen atom from a molecule of geraniol. Epoxygeraniol (**5**, corresponding to model compound **17**) has been detected in the autoxidation mixture.[7]

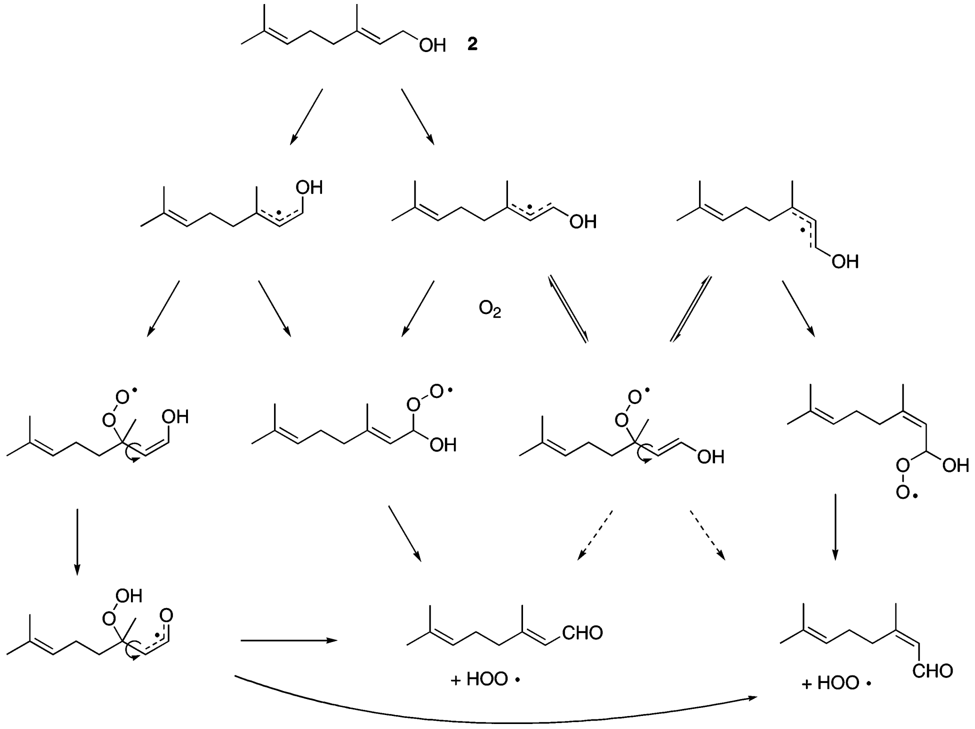


Figure 9. Expected open-shell autoxidation pathways starting from geraniol, **2**.

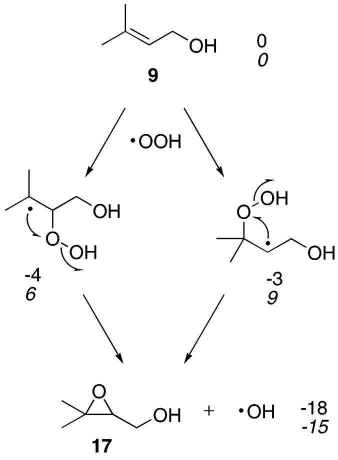


Figure 10. Epoxidation of geraniol model **9** by hydroperoxyl radical, with enthalpies and free energies (both at 298 K, the latter in italic) for all species.

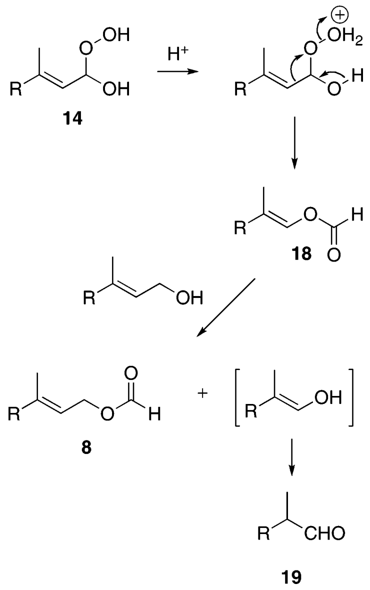


Figure 11. Formation of **8** through Baeyer−Villiger rearrangement and transesterification.

Finally, we shall discuss the formation of geranyl formate (**8**). This product differs from the other oxidation products in that it contains an additional carbon atom, which must have come from the fragmentation of another molecule of geraniol. We have not located any radical process leading to formates, but instead we speculate that it can be formed from perhydrate **14**, which can be formed either directly in the radical chain process as depicted in Figure 4 or by the reversible addition of hydrogen peroxide to either geranial (**3**) or neral (**4**), all of which are present in the autoxidation mixture. We note that **14** is reminiscent of the text-book intermediate in the Baeyer−Villiger reaction. Under acidic conditions, **14** is expected to fragment by cleavage of the O−O bond with simultaneous migration of the vinyl moiety, forming a vinyl formate (**18**). The latter has not been detected, but under the slightly acidic conditions of the autoxidation mixture, it would be expected to transesterify irreversibly with a geraniol molecule, whereupon the produced enol would tautomerize to C_9 aldehyde **19** (Figure 11).

A weakness of the current proposal is that aldehyde **19**, or indeed any C_9 products, has so far not been identified in the autoxidation mixture. However, geraniol is the only source of carbon in the experiment, and thus the only possible precursor for the formate moiety in **8**. In a separate experiment, a sample of authentic aldehyde **19** was added to geraniol and subjected to the normal oxidation procedure, as described previously.[7] The concentration of **19** slowly decreased and could, after a while, not be detected anymore. As negative evidence, this should not be taken as mechanistic proof, but it at least indicates that the absence of **19** in the autoxidation mixture does not disprove the mechanism depicted in Figure 11. However, other Baeyer−Villiger-type mechanisms can also be proposed, since both hydrogen peroxide and hydroperoxides are present together with aldehydes in the autoxidation mixture.

Conclusions

The autoxidation products of the monoterpene geraniol (**2**) have been rationalized computationally by investigation of plausible radical chain reactions for a model system. Both propagation steps in the accepted mechanism, radical chain transfer and the addition of O_2, were found to be exergonic, in contrast to the recently investigated isomeric linalool system.[5] However, in addition to the normal chain transfer mechanism, the geraniol-derived peroxyl radicals can also undergo intramolecular hydrogen abstraction followed by fragmentation, liberating a hydroperoxyl radical as an alternative chain transfer agent. The latter process was found to be favored compared to the classical intermolecular hydrogen abstraction. Either process gives as a side product the observed hydrogen peroxide. Some of the located intermediates allow cis−trans isomerization of the original geraniol double bond, rationalizing the observation of both geranial and neral as oxidation products.

Secondary oxidation products like epoxides and formates were also considered, and plausible reaction pathways for the formation of both have been advanced, in the former case based on the oxidation of geraniol by a hydroperoxyl radical, in the latter case through a Baeyer−Villiger rearrangement of one of the oxidation intermediates.

Acknowledgment. The work of L.H. was financially supported by the Research Institute for Fragrance Materials, Inc. (RIFM). Support from the Swedish Research Council is gratefully acknowledged. The work is part of the platform Göteborg Science Center for Molecular Skin Research at the Faculty of Science, Göteborg University.

References

(1) Karlberg, A.-T.; Magnusson, K.; Nilsson, U. *Contact Dermatitis* **1992**, *26*, 332−340.

(2) Karlberg, A.-T.; Shao, L. P.; Nilsson, U.; Gäfvert, E.; Nilsson, J. L. G. *Arch. Derm. Res.* **1994**, *286*, 97−103.

(3) Sköld, M.; Börje, A.; Matura, M.; Karlberg, A.-T. *Contact Dermatitis* **2002**, *46*, 267−272.

(4) Sköld, M.; Börje, A.; Harambasic, E.; Karlberg, A.-T. *Chem. Res. Toxicol.* **2004**, *17*, 1697−1705.

(5) Bäcktorp, C.; Johnson Wass, J. R. T.; Panas, I.; Sköld, M.; Börje, A.; Nyman, G. *J. Phys. Chem. A* **2006**, *110*, 12204−12212.

(6) Wang, G.; Zhang, D.; Xu, X.; Zhou, J. *J. Phys. Chem. A* **2007**, *111*, 747−752.

(7) Hagvall, L.; Bäcktorp, C.; Svensson, S.; Nyman, G.; Börje A.; Karlberg, A.-T. *Chem. Res. Toxicol.* **2007**, *20*, 809−814.

(8) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785. (c) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(9) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian 03*, Revision B.05; Gaussian, Inc.: Wallingford, CT, 2004.

(10) (a) Olivella, S.; Solé, A. *J. Am. Chem. Soc.* **2003**, *125*, 10641−10650. (b) Pratt, D. A.; Mills, J. H.; Porter. N. A. *J. Am. Chem. Soc.* **2003**, *125*, 11827−11828.

(11) Rienstra-Kiracofe, J. C.; Allen, W. D.; Schaefer, H. F., III. *J. Phys. Chem. A* **2000**, *104*, 9823−9840.

CT7001495

JCTC Journal of Chemical Theory and Computation

CHARMM Force Field Parameters for Nitroalkanes and Nitroarenes

Jeffery B. Klauda* and Bernard R. Brooks

Laboratory of Computation Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892

Received August 1, 2007

Abstract: New CHARMM force field (FF) parameters are developed for nitro compounds, referred to here as C27rn, for subsequent use in molecular dynamics (MD) simulations. The nonbonded terms are adjusted to best fit densities and hydration energies of nitropropane and nitrobenzene. High-level quantum mechanical calculations are used to obtain accurate conformational energies of nitroalkanes and nitrobenzene and to adjust the torsional potential of the CHARMM FF. For nitroalkanes, the calculated gauche (*g*) conformer of the C−C−C−N torsion is more stable than trans (*t*). Consequently, nitropropane MD simulations with C27rn result in 74% population of this *g* conformer. The C27rn FF is in excellent agreement with experiment for various bulk (density, isothermal compressibility, and heat of vaporization) and interfacial (surface tension) properties of nitropropane, nitrobutane, and nitrobenzene. MD simulations with the OPLS-AA FF for nitropropane and nitrobenzene result in similar property predictions as C27rn, except a reduced stability of the C−C−C−N *g* conformer.

1. Introduction

Compounds containing one or more nitro groups are commonly used as explosives,[1] organic solvents,[2,3] herbicides,[4] pesticides,[4] and drugs.[5,6] A few specific examples of these nitro compounds are described briefly as motivation for their general importance. In 1947, the first broad spectrum antibiotic (chloramphenicol) was discovered.[6] This early antibiotic contains a nitro group attached to a benzene ring but is not used extensively because of bacterial resistance and certain undesirable side effects. As another example, pure nitrobenzene or 2-nitrophenyl *n*-octyl ether are widely used as organic solvents.[2,3] These compounds are prominent in studies of ion transfer across interfaces with two immiscible liquids. The fluorescence of nitrobenzene with tryptophan has been important in binding studies of substrates in proteins. Specifically, sugar binding studies of the transmembrane protein lactose permease have used intrinsic Trp fluorescence with a nitro containing sugar, 6′-(*N*-dansyl)-aminohexyl-1-thio-*β*-D-galactopyranoside (α-NPG), to determine and quantify sugar binding.[7,8]

Although nitro compounds are of general and biological significance, only a limited number of studies have focused on developing force field (FF) parameters for use in molecular simulations.[9−13] Price et al.[13] developed nitro parameters for the OPLS-AA FF that resulted in good agreement with experimental gas-phase and liquid properties, e.g., density, heats of vaporization, and free energies of salvation, from molecular simulations. The primary focus of FF development has been nitrobenzene because of its importance as an organic solvent. An excellent comparison of nitrobenzene FFs (OPLS-AA,[13] Michael and Benjamin,[11] and Janssen et al.[9]) and experiment is discussed by Jorge et al.[10] The FF by Michael and Benjamin[11] focused on the nitrobenzene/water interface, while other FFs were only compared with bulk properties. It was found that OPLS-AA compares most favorably with experiment[10] and will be used as a benchmark in our studies.

For the CHARMM FF, a parameter set is not currently available for nitro compounds.[14] Therefore, the main purpose of this work is to develop nitro parameters consistent with CHARMM optimization procedures.[14−16] This new parameter

* Corresponding author e-mail: jbklauda@umd.edu. Current address: Department of Chemical and Biomolecular Engineering, University of Maryland, College Park, MD 20742.

10.1021/ct700191v This article not subject to U.S. Copyright. Published 2008 by the American Chemical Society
Published on Web 12/06/2007

set is then tested on pure liquid and interfacial systems of nitroalkanes and nitrobenzene.

Special focus in the force field development is on potential energy scans of two nitro torsional angles, i.e., C−C−N−O and C−C−C−N. Several conformational energies of nitroalkanes with the OPLS-AA FF were compared with ab initio energies at the HF/6-31g(d) level.[13] For nitroaromatics, Staikova and Cszmadia[17] used the same quantum mechanical (QM) methods to study the conformational energies about the C−C−N−O torsion. However, we have demonstrated with alkanes the importance of including electron correlation, i.e., more accurate QM methods, for torsional energies.[18,19] Consequently, highly accurate ab initio methods will be used in this study to describe the conformational energies of nitroalkanes and nitrobenzene. The FF will be adjusted accordingly to best match these QM calculations, and the methods and results will be described in the following sections.

2. Methodology

The methodologies used for fitting the CHARMM FF (2.1), ab initio calculations (2.2), and molecular dynamics simulations (2.3) for nitro compounds are described in this section.

2.1. Force Field Fitting. The potential energy $V(\hat{R})$ in the CHARMM FF[14] as well as other additive FF[13,20−22] is a function of the positions of all of the atoms in the system and has the following general form:

$$V(\hat{R}) = \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \left[\sum_j K_{\varphi,j}(1 + \cos(n_j\varphi - \delta_j))\right] + \sum_{\text{nonbonded pairs}} \epsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \sum_{\text{nonbonded pairs}} \frac{q_i q_j}{\epsilon_D r_{ij}} \quad (1)$$

Urey-Bradley and improper dihedral terms are available in CHARMM but are not used for the nitro compounds. The parameters in the initial two intramolecular terms of eq 1 (K_b, b_0, K_θ, and θ_0) are obtained from previous fits for the OPLS-AA FF.[13] Identical methods are used in developing these force field parameters for CHARMM.[14,23] The dihedral potential is optimized based on highly accurate torsional energy scans from QM calculations on model compounds.[18,24,25] van der Waals interactions are treated by the well-known Lennard-Jones (LJ) "6-12" potential, where ϵ_{ij} is the potential energy minimum between two particles, and $R_{\min,ij}$ is the position of this minimum. The conversion factor between $R_{\min,ij}$ and σ_{ij} in other LJ potential forms is $R_{\min,ij} = 2^{1/6}\sigma_{ij}$. Last, q_i and q_j are the atomic partial charges, and ϵ_D is the dielectric constant.

The optimization procedure for the parameters in eq 1 is consistent with the procedure used for developing the CHARMM FF.[14−16] Parameters for only nitrogen, oxygen, and adjacent group atoms (carbons bonded to nitrogen) were adjusted to maintain the transferability of other atoms in the molecule. First, the atomic charges on the nitro compound are adjusted to best represent scaled QM calculations (details in section 2.2) of water/nitro compound interaction energies based on the TIP3P[26,27] water model. This is known as the supramolecule approach,[14] and the initial guess for the atomic charges is based on similar groups in the CHARMM FF. The LJ parameters are modified to best represent the experimental density and optimized separately for nitroalkanes and nitrobenzene to represent changes in the nitro dispersion energies due to the neighboring aromatic ring. Only the experimental density at 298.15 K for nitropropane and nitrobenzene is used for the LJ fits. Therefore, other properties, compounds, and temperatures are predictions. The dihedral parameters ($K_{\phi,j}$, n_j, and δ_j) are fitted to accurate QM conformational energies (details in section 2.2) and consist of 1−4 sets, j, per dihedral type and summed over all the dihedral angles in the molecule.

The optimization of the C27rn FF parameters is typically an iterative process and requires several changes to obtain proper convergence of the desired properties. The supermolecule approach defines the charges, but the LJ and dihedral parameters are interdependent. Changes to the LJ terms are made until the bulk density is in satisfactory agreement with experiment. Consequently, the dihedral parameters are optimized for each LJ parameter set.

2.2. Quantum Mechanical Calculations. The Gaussian03 suite of programs[28] was used for the following QM calculations: (1) conformational states of nitropropane, nitrobutane, nitropentane, and nitrobenzene and (2) water interacting with individual nitro compounds. The conformational minima were optimized using tight convergence criteria (1.5×10^{-4} and 1.0×10^{-4} hartree/bohr for maximum and rms force) and a starting structure near the corresponding conformation, i.e., trans or gauche. In addition to the minima, conformations between the local minima and barriers are optimized. This was done using the Berny Algorithm[29] by fixing a corresponding dihedral angle on a transition state (TS) pathway. Geometry optimizations for single molecule conformations were performed at MP2/cc-pVDZ, while HF/6-31g(d) was used for water/nitro interactions. This lower level of theory for the dimer was used to be consistent with the CHARMM parametrization of the Coulombic terms.[14]

The HM-IE method[30] was used to estimate the energy of each individual molecular conformation of nitro compounds at the CCSD(T) level with a basis set larger than that used to obtain the optimized geometry. This method estimates molecular properties by assuming that the separate effects of electron correlation and basis set size are additive. These hybrid or compound QM methods, such as the Gaussian-3,[31,32] Dunning and Peterson,[33] and HM-IE,[30] estimate energies of CCSD(T) with a large basis set (LBS) by calculating CCSD(T) with a smaller basis set (SBS) and adding a correction based on the difference between MP2 energies with a LBS and a SBS as follows

$$E^{\text{conf}}[\text{CCSD(T)/LBS}] = E^{\text{conf}}[\text{CCSD(T)/SBS}] + (E^{\text{conf}}[\text{CCSD(T)/LBS}] - E^{\text{conf}}[\text{CCSD(T)/SBS}]) \cong E^{\text{conf}}[\text{CCSD(T)/SBS}] + (E^{\text{conf}}[\text{MP2/LBS}] - E^{\text{conf}}[\text{MP2/SBS}])$$

$$\equiv E^{\text{conf}}[\text{MP2:CC}] \qquad (2)$$

where E^{conf} is the energy of the conformer, and the difference between CCSD(T)/LBS and CCSD(T)/SBS is approximated at the MP2 level. A SBS of cc-pVDZ and LBS of cc-pVQZ was used here and also found previously to be an accurate measure of CCSD(T)/cc-pVQZ for linear alkanes.[18]

As an additional accuracy measure of the dihedral potential, the curvatures for selected trans or gauche wells were estimated by taking the second derivative of the energy, $V(\phi)$, with respect to the angle by means of fitting a parabolic function

$$V(\phi) = k(\phi - \phi_0)^2 + l(\phi - \phi_0) + m \qquad (3)$$

where k, l, and m are fit to either the QM energy or empirical force field predictions with energies up to 2 kcal/mol higher than the local minima.

2.3. Molecular Dynamics Simulations. Simulations were performed with CHARMM[34] using C27r for the alkane portion of the FF[18,19] and adjustments to the nitro force field, referred to here as C27rn. Nitrobenzene and nitropropane simulations with OPLS-AA are also performed for comparison with C27rn. The leapfrog Verlet algorithm was used with cubic periodic boundary conditions. A time step of 1 fs was applied to ensure time step artifacts did not affect our calculated properties. LJ interactions were smoothed by a switching function over 8−10 Å. Isobaric−isothermal ensemble (NPT) simulations were run with long-range electrostatics and LJ corrections. The particle mesh Ewald (PME)[35] method was used for the long-range electrostatic contribution (beyond 10 Å) to the total energy with $\kappa = 0.34$ Å^{-1} and a fast-Fourier grid density of about 1 Å^{-1}. The isotropic periodic sum (IPS) method[36] was used to obtain the long-range correction in LJ at an effective infinite cutoff. This PME/IPS method has been found to be accurate in bulk and interfacial systems.[37] All hydrogen atoms were constrained using the SHAKE algorithm.[38] The extended system formalism was used to maintain the temperature via the Hoover thermostat[39] and/or pressure[40,41] with a thermostat coupling constant of 20 000 kcal mol^{-1} ps^{-2} and a piston mass of 2000 amu.

Initial conformations for bulk nitropropane, nitrobutane, and nitrobenzene were obtained by placing 320, 256, and 200 molecules, respectively, on an even grid in random orientations. With these starting conformations, the energy was minimized with the steepest descent routine for 200 steps to reduce unfavorable van der Waals contacts. The velocities were then set to the desired temperature, and an equilibration period of 500 ps was used for all simulations to ensure full equilibration. The coordinates were saved every 1 ps for a total simulation time of 2 ns for simulations at 288.15, 293.15, and 303.15 K, but a simulation time of 5 ns was used for the 298.15 K runs. For vapor simulations, N simulations with a single molecule were run for 500 ps after 50 ps of equilibration. Coordinates from the end of each liquid simulation at 298.15 K were used as N initial coordinates for the vapor simulations.

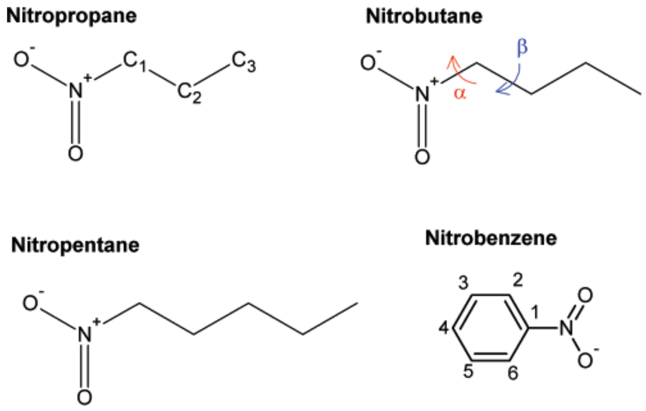


Figure 1. Model compounds used in QM calculations to develop the C27rn FF. The atom types for the aliphatic carbons are labeled on nitropropane, and the dihedrals are labeled on nitrobutane.

Table 1. Nonbonded Parameters for C27rn[a]

atom	description/location	q [e]	ϵ [kcal/mol]	$R_{\min}/2$ [Å]
N	nitroalkane	+0.50	−0.160	1.837
O	nitroalkane	−0.40	−0.120	1.700
C	C_1 in nitroalkane	+0.16	−0.056	2.010
H	attached to C_1	+0.07	−0.028	1.340
N	nitroarene	+0.50	−0.120	1.850
O	nitroarene	−0.40	−0.100	1.770
C	C_6 in nitrobenzene	+0.34	−0.070	1.992
C	C_1 or C_5 in nitrobenzene	−0.18	−0.070	1.992
H	attached to C_1 or C_5	+0.16	−0.046	1.100

[a] See Figure 1 for labeling nomenclature.

Table 2. Torsional Parameters for C27rn[a]

	K_ϕ [kcal/mol]	n	δ [deg]
CH2−CH2−N−O	0.060	2	0
CH3(2)−CH2−CH2−N	0.084	4	0
	0.360	3	0
	0.151	2	0
	0.133	1	180
HA−CA−CA−N	1.000	2	180
CA−CA−CA−N	6.140	2	180
CA−CA−N−O	1.100	2	180

[a] CH3 and CH2 are for nitroalkanes and CA is for nitrobenzene.

Densities, heat of vaporization, isothermal compressibilities, and self-diffusivities were calculated. Standard errors were estimated from block averages.[42] Isothermal compressibilities were calculated from

$$\beta_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T = \frac{\langle \delta V^2 \rangle}{V k_b T} \qquad (4)$$

where V is the volume, $\langle \delta V^2 \rangle$ is the volume fluctuation, and k_b is Boltzmann's constant. The 5-ns simulations at 298.15 K were used to calculate eq 4. The slope of the mean squared displacement versus time was used to determine the apparent self-diffusivity for the periodic boundary condition, D_{PBC}, using a weighted least squared fit with weights obtained from averages of 8−10 subgroups of molecules per trajectory. The self-diffusivity was corrected for system-size effects using the hydrodynamic model of Yeh

Table 3. Molecular Structures of Nitro Compounds from MP2/cc-pVDZ (except Nitropropane Also with cc-pVTZ) with Atom Numbering and Dihedral Angles as Shown in Figure 1[a]

	nitropropane				nitrobutane		nitropentane		nitrobenzene	
	t-DZ	*g*-DZ	*t*-TZ	*g*-TZ	*t*	*g*	*t*	*g*	MP2	exp[b]
C_1-N	1.50	1.50	1.49	1.49	1.50	1.50	1.50	1.50	1.48	1.49
C_1-C_2	1.52	1.53	1.52	1.52	1.52	1.53	1.52	1.53	1.40	1.40
$N{=}O$	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.22
$\angle C_1-N{=}O$	117.1	117.0	117.1	117.2	117.1	117.0	117.1	117.0	117.3	117.3
$\angle O{=}N{=}O$	125.9	125.9	125.6	125.5	125.9	125.9	125.9	125.9	125.4	125.3
$\angle C_1-C_2-N$	110.6	109.1	109.2	109.4	110.5	109.0	110.6	109.0	118.7	118.3
α	121.7	111.1	88.6	115.1	121.6	110.2	121.9	110.1	0.0	
	−57.9	−66.7	−88.6	−63.3	−58.0	−67.5	−57.7	−67.6		
β	179.7	−60.7	180.0	−59.0	180.0	−61.1	179.6	−60.9	0.0	

[a] Distances are in Å and angles are in degrees. For the nitroalkanes, structures are listed in the (CCCN) trans (*t*) and gauche (*g*) conformation. [b] The experimental electron diffraction data for nitrobenzene.[44,45]

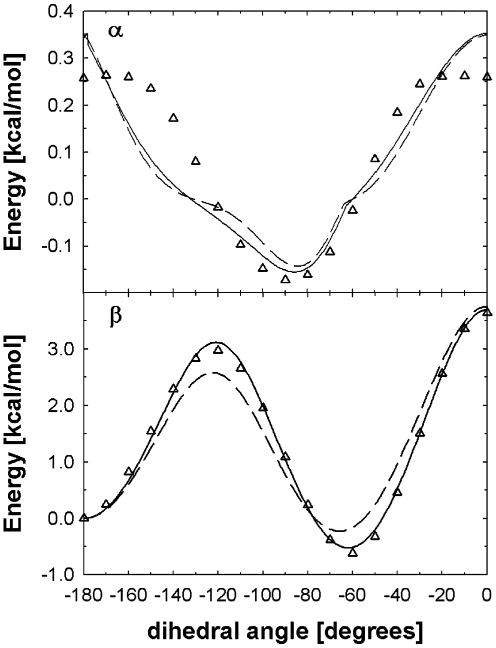


Figure 2. Conformational energies of nitropentane as a function of α (C_2-C_1-N-O) and *β* ($C_3-C_2-C_1-N$) torsional angles. The corresponding dihedral is fixed for each point on the panel, but all other degrees of freedom are minimized. The symbols are QM energies (MP2:CC), the solid line is C27rn, and the dashed line is OPLS-AA.

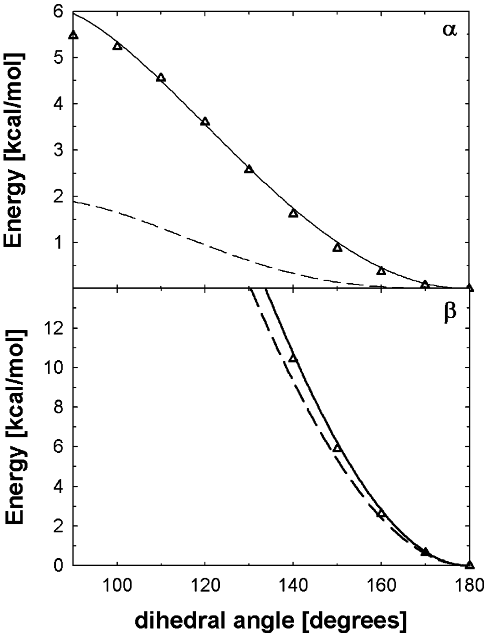


Figure 3. Conformational energies of nitrobenzene as a function of α and *β* torsional angles. The corresponding dihedral is fixed for each point on the panel, but all other degrees of freedom are minimized. The symbols are QM energies (MP2:CC), the solid line is C27rn, and the dashed line is OPLS-AA.

and Hummer[43] of a particle surrounded by a solvent with a viscosity, η

$$D_s = D_{PBC} + \frac{k_B T \xi}{6\pi\eta L} \tag{5}$$

where L is the cubic box length and $\xi = 2.837297$.[43]

The heat of vaporization was calculated from

$$\Delta H^{vap} = \langle U_{ig}\rangle - \frac{\langle U_l\rangle}{N} + RT \tag{6}$$

where $\langle U_l\rangle$ is the average internal energy over time (sum of intra- and intermolecular energies) of the liquid state, N is the total number of molecules, and $\langle U_{ig}\rangle$ is the average ideal gas internal energy. The average liquid internal energy was obtained from the liquid simulations and N gas simulations to obtain $\langle U_{ig}\rangle$.

The surface tension was evaluated from

$$\gamma = 0.5\langle L_z[P_{zz} - 0.5(P_{xx} + P_{yy})]\rangle \tag{7}$$

where L_z is the size of the simulation box normal to the interface, P_{zz} is the normal component of the internal pressure tensor, and P_{xx} and P_{yy} are the tangential components. The MD simulations here contain two interfaces (a liquid film with vapor at the top and bottom, see ref 37), so a prefactor of 0.5 is required to obtain γ on a per interface basis.

3. Results and Discussion

The parametrization of the LJ, electrostatics, and dihedral terms is iterative, but the results discussed here are based on the optimal values in Tables 1 and 2. The ab initio calculations on the molecular structure and torsional profiles are presented first. Then, the conformational energies of the

Table 4. Nitroalkane Conformer Energies in kcal/mol Relative to the All-Trans State of the β Torsion[a]

		ΔE_g	$\Delta E^{\dagger}_{t/g}$	$\Delta E^{\dagger}_{g/g}$
C_3NO_2	MP2:CC	−0.60	3.14	3.68
	C27rn	−0.49	3.13	3.64
	OPLS-AA	0.06	2.64	4.04
		ΔE_g	$\Delta E^{\dagger}_{t/g}$	$\Delta E^{\dagger}_{g/g}$
C_4NO_2	MP2:CC	−0.61	2.96	3.65
	C27rn	−0.56	3.10	3.65
	OPLS-AA	−0.24	2.57	3.73
		ΔE_g	$\Delta E^{\dagger}_{t/g}$	$\Delta E^{\dagger}_{g/g}$
C_5NO_2	MP2:CC	−0.62	2.98	3.64
	C27rn	−0.52	3.12	3.70
	OPLS-AA	−0.23	2.58	3.76

[a] MP2:CC is the approximate CCSD(T)/cc-pVQZ energy using eq 2. C27rn is the modified C27r force field. The energy of the transition state between local minima *i* and *j* relative to the all-trans state is denoted as $\Delta E^{\dagger}_{i/j}$.

new force field (C27rn) are compared with the QM calculations as well as the OPLS-AA.[13] Finally, C27rn is tested with bulk and interfacial simulations and compared with experiment and OPLS-AA.

3.1. Molecular Structures and Torsional Profiles. *3.1.1. Ab Initio Calculations.* The geometry of the nitro compounds (Figure 1) is optimized using MP2, and the distances, angles, and torsional angles are listed in Table 3. The agreement between electron diffraction[44,45] and MP2/cc-pVDZ for nitrobenzene is excellent, i.e., deviations of ≤0.01 Å and 0.4°. For nitroalkanes, there are two minima of the C−C−C−N torsion (β), i.e., *t*-trans and *g*-gauche (g^- is shown here but equivalent to g^+). Double- and triple-ζ basis set optimizations for the *t* conformation of nitropropane result in a noticeable difference for the C−C−N−O torsion angle, α (Table 3). However, in terms of conformational energy (of most interest to this work) HM-IE corrects for this basis set effect. Moreover, there are negligible differences with other structural properties of the *t* conformer and all properties of the *g* conformer. Therefore, MP2/cc-pVDZ optimizations will be used for their efficiency and reasonable accuracy with larger nitroalkanes. Previous calculations with B3LYP/6-31g(d)[46] result in *t* structures similar to those in Table 3 but may result in similar basis set problems. There are other structural differences between the *t* and *g* conformation, i.e., $\angle C_1-C_2-N$ and α are reduced in the *g* conformation compared to *t*. However, other internal geometry values are not influenced by this change in the β torsion.

Quantum mechanical conformational energies of the α and β torsions of the four nitro compounds were calculated using eq 2 with a total of 129 conformations. The torsional profiles for nitropentane and nitrobenzene are shown as examples in Figures 2 and 3, respectively. The β torsion is alkane-like with two minima (*t* and *g*). For alkanes, high-level ab initio QM calculations on pentane through heptane yield a ΔE_g (energy difference from the all-trans state) slightly higher than +0.5 kcal/mol.[18,47−49] However, the *g* conformation in nitroalkanes is lower in energy, i.e., $\Delta E_g = -0.6$ kcal/mol (see Table 4). The terminal nitrogen and oxygen on this β torsion stabilize the *g* state. Similarly, for nitroalkanes there is a greater than 1 kcal/mol decrease in the cis conformational energy compared to alkanes. Although there are differences in *g* and cis energies, the conformational energy of the transition from the *t* to *g* state $\Delta E^{\dagger}_{t/g}$ is similar to that of an alkane (3 kcal/mol).[18]

The conformational energy barriers for nitroalkanes are lower than nitrobenzene. The conformational space is restricted because the nitro group is attached to an aromatic ring (Figure 3). Therefore, the lowest energy conformation of nitrobenzene is when the nitrogen and oxygen atoms are in the same plane as the carbons, i.e., a planar molecule.

3.1.2. Empirical Potentials. A root-mean squared error objective function was used to fit the set of nitro dihedrals to the high-level QM energies discussed above. Table 2 lists five sets of dihedrals fit to the ab initio calculations, denoted here as C27rn. Only the C−C−C−N dihedral required more than one term. This is similar to the alkane C−C−C−C torsion in the C27r FF,[18] where multiple torsional terms were needed to accurately fit conformational energies.

The molecular structure of the nitroalkanes and nitrobenzene with C27rn is listed in Table 5. The bond lengths are nearly identical between the MP2/cc-pVDZ and C27rn optimized structures. $\angle C_1-C_2-N$ is slightly larger with C27rn but only deviates by 3−4°. The optimized *t* conformation with C27rn results in an α torsion in good agreement with the correct value using MP2/cc-pVTZ. There is also excellent agreement for the value of the β torsion of the *g* conformation with less than 1° difference between QM and C27rn.

The calculated nitroalkane minima and transition state (TS) energies with C27rn and OPLS-AA are compared with MP2:CC in Table 4. The absolute average deviation (AAD) from MP2:CC for these conformations with C27rn and OPLS-AA is 0.07 and 0.36 kcal/mol, respectively. The AAD for just ΔE_g is larger with OPLS-AA (0.47 kcal/mol) and similar for C27rn (0.09 kcal/mol). Overall, C27rn is superior to OPLS-AA in these conformational energies. OPLS-AA is parametrized on smaller molecule conformations of nitroalkanes with HF/6-31g(d). Since low-level QM calculations are known to result in inaccurate dispersion energies,[50,51] the large discrepancy in ΔE_g is not surprising for OPLS-AA.

The correct curvature of *t* and *g* wells is of greater importance than slight inaccuracies of minima energies, because there is an increased availability of dihedral angles at a given temperature.[18] Table 6 lists the curvature for the minima of the *t* and *g* conformers of the β torsion calculated from eq 3. The curvature of the C27rn wells compared to OPLS-AA is in better agreement with MP2:CC with an overall AAD of 0.16×10^{-3} and 0.63×10^{-3} kcal mol^{-1} deg^{-2}, respectively. The QM curvature of the *g* well is greater than *t* due to electron repulsion. For OPLS-AA the *g* well is too broad (Table 6) because of the lack of conformations other than the minima and transitional barrier and a less accurate QM method.[13]

OPLS-AA and C27rn both follow the conformational energies of MP2:CC qualitatively as shown in Figures 2 and 3, but C27rn is noticeably in better agreement with QM. Although there are differences with QM and C27rn for the

Table 5. Molecular Structures of Nitro Compounds with C27rn

	nitropropane		nitrobutane		nitropentane		nitrobenzene	
	t	*g*	*t*	*g*	*t*	*g*	C27rn	exp[a]
C_1-N	1.50	1.50	1.50	1.50	1.50	1.50	1.47	1.49
C_1-C_2	1.54	1.54	1.54	1.54	1.54	1.54	1.42	1.40
$N{=}O$	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.22
$\angle C_1-N{=}O$	117.5	117.4	117.5	117.5	117.5	117.5	117.5	117.3
$\angle O{=}N{=}O$	124.5	124.2	124.6	124.4	124.5	124.4	125.0	125.3
$\angle C_1-C_2-N$	113.8	113.8	113.7	113.9	113.7	113.9	121.7	118.3
α	85.8	115.2	86.4	108.7	85.9	110.1	0.0	
	−85.5	−75.4	−86.2	−78.6	−86.0	−77.8		
β	180.0	−61.4	180.0	−61.9	180.0	−61.9	0.0	

[a] The experimental electron diffraction data for nitrobenzene.[44,45] Atom numbering and dihedral angles as shown in Figure 1.

Table 6. Curvatures (2*k* in Eq 3) of Trans And Gauche Conformations[a]

state	molecule	MP2:CC	C27rn	OPLS-AA
t	C_3NO_2	2.57	2.61	2.26
g	C_3NO_2	4.26	3.98	2.99
t	C_4NO_2	2.42	2.60	2.23
g	C_4NO_2	4.14	3.98	3.16
t	C_5NO_2	2.42	2.61	2.23
g	C_5NO_2	4.18	3.98	3.16

[a] In units of 10^{-3} kcal mol^{-1} deg^{-2}.

Table 7. C27rn Simulation Averages and Standard Errors for Dipole Moment (μ), Density (ρ), Isothermal Compressibility (β_T), Diffusivities (D_{PBC} and D_s), and Heat of Vaporization (ΔH^{vap}) at 298.15 K[a]

		C_3NO_2	C_4NO_2	NB
μ	C27rn	4.80 ± 0.00	4.80 ± 0.00	4.53 ± 0.00
[D]	OPLS-AA	3.82 ± 0.00		3.33 ± 0.00
	exp[53]	3.59		4.22
ρ	C27rn	0.999 ± 0.007	0.974 ± 0.007	1.208 ± 0.009
[g/cm^3]	OPLS-AA	0.974 ± 0.008		1.154 ± 0.009
	exp[52]	0.996	0.968	1.198
β_T	C27rn	6.27 ± 0.15	4.64 ± 0.16	4.17 ± 0.10
[10^{-10} m^2/N]	OPLS-AA	11.34 ± 0.09		4.57 ± 0.12
	exp[53]			5.23
D_{PBC}	C27rn	0.712 ± 0.070	0.595 ± 0.010	0.370 ± 0.060
[10^{-5} cm^2/s]	OPLS-AA	0.580 ± 0.083		0.706 ± 0.072
	exp[53]			1.08
D_s	C27rn	0.929 ± 0.070	0.815 ± 0.010	0.477 ± 0.060
[10^{-5} cm^2/s]	OPLS-AA	0.792 ± 0.083		0.812 ± 0.072
	exp[53]			1.08
ΔH^{vap}	C27rn	12.72 ± 0.38		14.19 ± 0.60
[kcal/mol]	OPLS-AA			13.05 ± 0.48
	exp[53]	10.37		13.15

[a] β_T and ΔH^{vap} were calculated from eqs 4 and 6, respectively. D_{PBC} is the apparent self-diffusivity obtained directly from the mean squared displacement in the simulations, and the corrected self-diffusivity, D_s, is obtained from eq 5.

Table 8. Surface Tension of Nitropropane and Nitrobenzene in dyn/cm Compared with Experiment[53]

		C_3NO_2	NB
293.15 K	C27rn	31.51 ± 0.87	43.52 ± 0.38
	exp	30.64	42.70
303.15 K	C27rn	29.49 ± 0.35	41.94 ± 0.64
	exp	29.61	42.17

α torsion in nitroalkanes (Figure 2), they are small for an essentially freely rotating potential at 298.15 K. Since the cause for this discrepancy is the LJ and electrostatic energies, a polarizable FF may improve this agreement between QM and C27rn. For the β torsion, the TS energy is lower than QM with OPLS-AA, and the *g* minimum is skewed for nitropentane (Figure 2), where as C27rn follows the QM energies almost exactly. The α adiabatic surface with OPLS-AA and C27rn is similar and in satisfactory agreement with MP2:CC for nitroalkanes. However, OPLS-AA significantly underpredicts the out-of-plane energy of the oxygen in nitrobenzene (>3 kcal/mol), where C27rn results in excellent agreement with MP2:CC.

3.2. Molecular Dynamics Simulations. The validity of the new C27rn FF for nitro compounds was tested on bulk and interfacial systems containing nitropropane, nitrobutane, and nitrobenzene. The nitrogen and oxygen LJ parameters for nitroalkanes were adjusted to best fit the density of nitropropane at 298.15 K, and the simulated density is in perfect agreement with experiment.[52] The temperature dependence of the density is shown in Figure 4, and the AAD of C27rn is only 0.29 g/cm^3. There is similar agreement for the densities of nitrobutane and nitrobenzene with an AAD of 0.53 and 0.85 g/cm^3, respectively (Figure 4). The LJ nitrogen and oxygen parameters for nitrobenzene were allowed to vary from nitroalkanes to obtain accurate densities at 298.15 K. The density of OPLS-AA for nitropropane and nitrobenzene are slightly lower than experiment (0.974 and 1.154 g/cm^3, respectively), but the LJ parameters for nitrogen and oxygen were identical for all nitro compounds in the OPLS-AA FF.

As shown previously, the conformational energies about the β torsional surface of nitroalkanes differ between OPLS-AA and C27rn. Consequently, the conformational probabilities for the β torsion are also quite different (Figure 5). There is an increased population of *g* conformations with C27rn (74% versus 57% for C27rn and OPLS-AA, respectively). This increase in *g* population is similar to MD simulations of pure alkanes[18,19] with a more accurate description of the C−C−C−C torsional surface. Contrary to nitroalkanes, experimental data are available for conformational populations of alkanes, and the C27r force field fit to QM (MP2:CC) is in excellent agreement with experiment.[18]

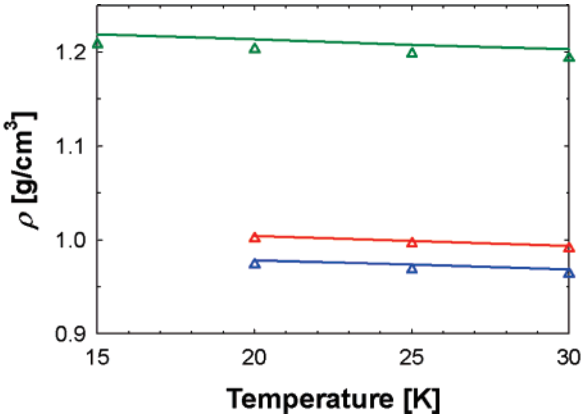


Figure 4. The density of nitrobenzene (green), nitropropane (red), and nitrobutane (blue) as a function of temperature for C27rn (lines) and experiment (triangles).

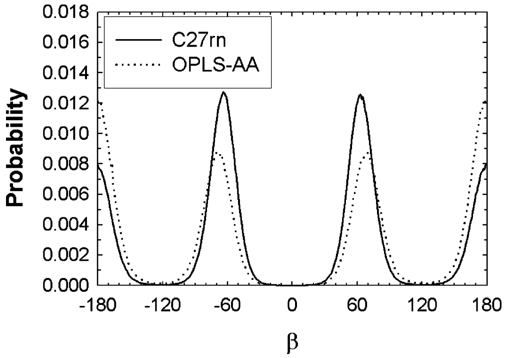


Figure 5. The probability of the β ($C_3{-}C_2{-}C_1{-}N$) torsional angle of nitropropane.

This implies that these results for nitro compounds using the same QM methods would be accurate if experimental data were available.

The molecular dipoles (μ) of nitropropane and nitrobenzene with C27rn are consistently higher than experiment.[53] An elevated μ is expected and typical of the CHARMM FF because the experimentally measured values are usually in the gas phase and polarization of the liquid phase results in an increase in the dipole moment. MD simulations with OPLS-AA result in an average μ of 3.82 and 3.33 D for nitropropane and nitrobenzene, respectively, which are significantly lower than experiment.

The radial distribution functions (RDFs) between oxygen and carbon calculated from the bulk simulations are shown in Figure 6 for nitrobenzene (top) and nitropropane (bottom). The RDFs for OPLS-AA and C27rn are similar in shape with some subtle differences in peak locations. For example, the O/C distance is slightly larger between molecules for OPLS-AA and consistent with an increase in the overall density of the C27rn liquids. There is a preference of oxygen to interact closely with the *para* (C4) and *meta* (C3 and C5) carbons in nitrobenzene (Figure 1, data not shown for *para*) but at a distance greater than a C−H···O hydrogen bond. However, interactions are weaker with the more negative *ortho* carbons (C2 and C6). A similar trend is seen for nitropropane with oxygen interacting more strongly with the C3 carbon. Moreover, the first peak height is reduced with OPLS-AA for the O/C1 and O/C3 RDFs.

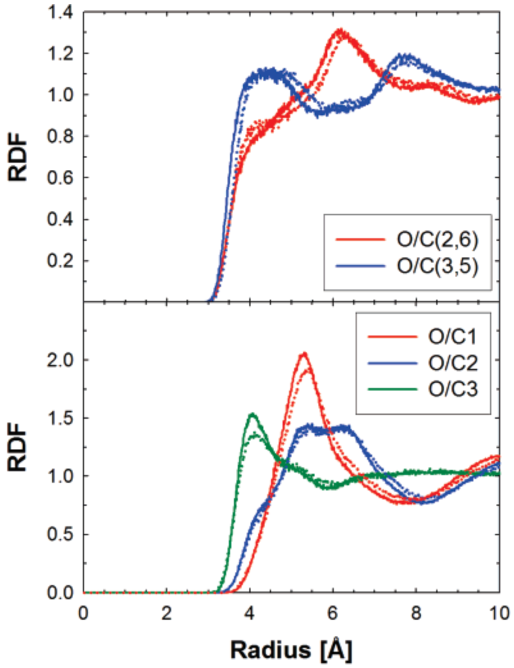


Figure 6. Radial distribution functions (RDF) between oxygen and carbon of nitrobenzene (top) and nitropropane (bottom). The C27rn results are in solid lines, and the OPLS-AA results are in dotted lines.

The enthalpy of vaporization (ΔH^{vap}) is comparable to experiment[53] for nitropropane and nitrobenzene (Table 7). C27rn results in ΔH^{vap} that is slightly larger than experiment for nitropropane but within statistical error for nitrobenzene. The OPLS-AA FF is in slightly better agreement with experiment (13.05 kcal/mol for nitrobenzene) but at the cost of a lower bulk density. The isothermal compressibilities (β_T) of C27rn and OPLS-AA are lower than experiment for nitrobenzene by 1.06 and 0.66 × 10^{-10} m^2/N, respectively.

The diffusion constant (D_s) and surface tension (γ) were used as additional measures for the accuracy of C27rn. D_s of nitrobenzene is smaller than experiment for C27rn and OPLS-AA (0.812 × 10^{-5} cm^2/s). The nitrobenzene diffusion constant is larger for OPLS-AA compared to C27rn, but the OPLS-AA D_s for nitropropane is 15% smaller than C27r. This flip-flop in D_s order may not be the result of inaccurate nitro parameters, rather differences in the parameters for the alkane or benzene portion of the respective FF. The agreement with experiment is improved for γ (Table 8) compared to D_s with an AAD of 1.6 and 1.2 dyn/cm for nitropropane and nitrobenzene, respectively.

4. Summary

Force field parameters for the important nitro group have been optimized for use with the CHARMM FF. The conformational energies of nitroalkanes and nitrobenzene were best fit to accurate and high-level QM calculations. Consequently, MD simulations with C27rn result in an increased population of *g* conformers compared to OPLS-AA. Bulk and interfacial properties from the nitro simulations with C27rn are in excellent agreement with experiment, especially densities, heats of vaporization, and surface tensions. However, the calculated diffusion constant of liquid nitrobenzene with both OPLS-AA and C27rn is lower than experiment. Simple nonpolarizable FF models can accurately model most liquid properties but without more complex functions, such as polarizability, not all parameters will be in excellent agreement with experiment. Since these new parameters accurately represent interaction energies between water and nitro compounds and pure component properties, C27rn can be used in simulations of biologically relevant compounds, such as antibiotics with nitro groups or sugar analogs as substrates in membrane proteins.

Acknowledgment. This research was supported in part by the Intramural Research Program of the NIH, National Heart, Lung and Blood Institute. Some of the simulation results utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). We would also like to thank Prof. Alexander MacKerell, Jr., for providing us with initial LJ and charge parameters for the nitro containing compounds.

References

(1) Spain, J. C. Biodegradation of Nitroaromatic Compounds. *Annu. Rev. Microbiol.* **1995**, *49*, 523.

(2) Marcus, Y. *Ion Properties*; Marcel-Dekker: New York, 1997.

(3) Scholz, F.; Schroder, U.; Gulaboski, R. *Electrochemistry of Immobilized Particles and Droplets*; Springer: Heidelberg, Berlin, 2005.

(4) Harrison, M. A. J.; Barra, S.; Borghesi, D.; Vione, D.; Arsene, C.; Olariu, R. L. Nitrated phenols in the atmosphere: a review. *Atmos. Environ.* **2005**, *39*, 231.

(5) Ahlner, J.; Andersson, R. G. G.; Torfgard, K.; Axelsson, K. L. Organic Nitrate Esters-Clinical Use And Mechanisms Of Actions. *Pharmacol. Rev.* **1991**, *43*, 351.

(6) Balbi, H. J. Chloramphenicol: A review. *Pediatr. Rev.* **2004**, *25*, 284.

(7) Nie, Y. L.; Smirnova, I.; Kasho, V.; Kaback, H. R. Energetics of ligand-induced conformational flexibility in the lactose permease of Escherichia coli. *J. Biol. Chem.* **2006**, *281*, 35779.

(8) Smirnova, I. N.; Kasho, V. N.; Kaback, H. R. Direct sugar binding to LacY measured by resonance energy transfer. *Biochemistry* **2006**, *45*, 15279.

(9) Janssen, R. H. C.; Theodorou, D. N.; Raptis, S.; Papadopoulos, M. G. Molecular simulation of static hyper-Rayleigh scattering: A calculation of the depolarization ratio and the local fields for liquid nitrobenzene. *J. Chem. Phys.* **1999**, *111*, 9711.

(10) Jorge, M.; Gulaboski, R.; Pereira, C. M.; Cordeiro, M. Molecular dynamics study of nitrobenzene and 2-nitrophenyloctyl ether saturated with water. *Mol. Phys.* **2006**, *104*, 3627.

(11) Michael, D.; Benjamin, I. Molecular dynamics simulation of the water|nitrobenzene interface. *J. Electroanal. Chem.* **1998**, *450*, 335.

(12) Price, D. J.; Brooks, C. L. Detailed considerations for a balanced and broadly applicable force field: A study of substituted benzenes modeled with OPLS-AA. *J. Comput. Chem.* **2005**, *26*, 1529.

(13) Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field. *J. Comput. Chem.* **2001**, *22*, 1340.

(14) MacKerell, A. D., Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584.

(15) MacKerell, A. D., Jr. Interatomic Potentials: Molecules. In *Handbooks of Material Modeling*; Yip, S., Ed.; Springer: The Netherlands, 2005; p 509.

(16) MacKerell, A. D., Jr. Atomistic Models and Force Fields. In *Computational Biochemistry and Biophysics*; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker: New York, 2001; p 7.

(17) Staikova, M.; Csizmadia, I. G. Ab initio investigation of internal rotation in conjugated molecules and the orientation of NO2 in nitroaromatics: nitrobenzene, o-monofluoro- and o,o′-difluoro-nitrobenzenes. *J. Mol. Struct.* (THEOCHEM) **1999**, *467*, 181.

(18) Klauda, J. B.; Brooks, B. R.; MacKerell, A. D., Jr.; Venable, R. M.; Pastor, R. W. An Ab Initio Study on the Torsional Surface of Alkanes and its Effect on Molecular Simulations of Alkanes and a DPPC Bilayer. *J. Phys. Chem. B* **2005**, *109*, 5300.

(19) Klauda, J. B.; Pastor, R. W.; Brooks, B. R. Adjacent gauche stabilization in linear alkanes: Implications for polymer models and conformational analysis. *J. Phys. Chem. B* **2005**, *109*, 15684.

(20) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(21) Jorgensen, W. L.; Maxwell, D. S.; Tiradorives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(22) Schuler, L. D.; Daura, X.; Van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205.

(23) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.

(24) Feig, M.; MacKerell, A. D., Jr.; Brooks, C. L. Force field influence on the observation of π-helical protein structures in molecular dynamics simulations. *J. Phys. Chem. B* **2003**, *107*, 2831.

(25) Woodcock, H.; Moran, D.; Pastor, R. W.; MacKerell, A. D., Jr.; Brooks, B. R. Ab initio modeling of glycosyl torsions and anomeric effects in a model carbohydrate: 2-Ethoxy Tetrahydrophyran. *Biophys. J.* **2007**, *93*, 1.

(26) Durell, S. R.; Brooks, B. R.; Bennaim, A. Solvent-Induced Forces between Two Hydrophilic Groups. *J. Phys. Chem.* **1994**, *98*, 2198.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.

(28) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.03) ed.*; Gaussian, Inc: Pittsburgh, PA, 2003.

(29) Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **1982**, *3*, 214.

(30) Klauda, J. B.; Garrison, S. L.; Jiang, J.; Arora, G.; Sandler, S. I. HM-IE: Quantum Chemical Hybrid Methods for Calculating Interaction Energies. *J. Phys. Chem. A* **2004**, *108*, 107.

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

(32) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) theory for molecules containing first and second-row atoms. *J. Chem. Phys.* **1998**, *109*, 7764.

(33) Dunning, T. H.; Peterson, K. A. Approximating the basis set dependence of coupled cluster calculations: Evaluation of perturbation theory approximations for stable molecules. *J. Chem. Phys.* **2000**, *113*, 7799.

(34) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM-a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(35) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald-an NLog(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(36) Wu, X. W.; Brooks, B. R. Isotropic periodic sum: A method for the calculation of long-range interactions. *J. Chem. Phys.* **2005**, *122*, 044107.

(37) Klauda, J. B.; Wu, X. W.; Pastor, R. W.; Brooks, B. R. Long-range Lennard-Jones and Electrostatic Interactions in Interfaces: Application of the Isotropic Periodic Sum Method. *J. Phys. Chem. B* **2007**, *111*, 4393.

(38) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Contraints: Molecular Dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327.

(39) Hoover, W. G. Canonical Dynamics-Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695.

(40) Nosé, S.; Klein, M. L. A Study of Solid and Liquid Carbon Tetrafluoride Using the Constant Pressure Molecular-Dynamics Technique. *J. Chem. Phys.* **1983**, *78*, 6928.

(41) Andersen, H. C. Molecular-Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

(42) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(43) Yeh, I. C.; Hummer, G. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B* **2004**, *108*, 15873.

(44) Borisenko, K. B.; Bock, C. W.; Hargittai, I. Intramolecular Hydrogen-Bonding And Molecular-Geometry Of 2-Nitrophenol From A Joint Gas-Phase Electron-Diffraction And Ab-Initio Molecular-Orbital Investigation. *J. Phys. Chem.* **1994**, *98*, 1442.

(45) Domenicano, A.; Schultz, G.; Hargittai, I.; Colapietro, M.; Portalone, G.; George, P.; Bock, C. W. Molecular Structure of Nitrobenzene in the Planar and Orthogonal Conformations. *Struct. Chem.* **1989**, *1*, 107.

(46) Shlyapochnikov, I. A.; Khrapkovskii, G. M.; Shamov, A. G. Structure and vibrational spectra of mononitroalkanes. *Russ. Chem. Bull.* **2002**, *51*, 940.

(47) Smith, G. D.; Jaffe, R. L. Quantum chemistry study of conformational energies and rotational energy barriers in n-alkanes. *J. Phys. Chem.* **1996**, *100*, 18718.

(48) Salam, A.; Deleuze, M. S. High-level theoretical study of the conformational equilibrium of n-pentane. *J. Chem. Phys.* **2002**, *116*, 1296.

(49) *Smithsonian Physical Tables*, 9th ed.; New York, 1954.

(50) Klauda, J. B.; Sandler, S. I. Ab Initio Intermolecular Potentials for Gas Hydrates and Their Predictions. *J. Phys. Chem. B* **2002**, *106*, 5722.

(51) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Publication, Inc.: Mineola, NY, 1996.

(52) Toops, E. E. Physical Properties Of Eight High-Purity Nitroparaffins. *J. Phys. Chem.* **1956**, *60*, 304.

(53) Riddick, J.; Bunger, W.; Sakano, T. *Techniques of Chemistry: Organic Solvents, Physical Properties and Methods of Purification*, 4th ed.; Wiley: New York, 1986.

CT700191V

JCTC Journal of Chemical Theory and Computation

P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation

Berk Hess*

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

Received August 9, 2007

Abstract: By removing the fastest degrees of freedom, constraints allow for an increase of the time step in molecular simulations. In the last decade parallel simulations have become commonplace. However, up till now efficient parallel constraint algorithms have not been used with domain decomposition. In this paper the parallel linear constraint solver (P-LINCS) is presented, which allows the constraining of all bonds in macromolecules. Additionally the energy conservation properties of (P-)LINCS are assessed in view of improvements in the accuracy of uncoupled angle constraints and integration in single precision.

I. Introduction

In classical molecular simulation methods, such as molecular dynamics (MD), the time step is limited by the fastest motions, which are bond oscillations. These oscillations have a relatively high frequency and low amplitude. By replacing at least the bond vibrations involving hydrogen atoms by holonomic constraints the time step in molecular simulations can be increased by roughly a factor of 4. Constraints are often considered a more faithful representation of the physical behavior of bond vibrations which are almost exclusively in their vibrational ground state.

Constraints can be added to the Hamiltonian using Lagrange multipliers. When time is discretized the linear equations for the Lagrange multipliers become nonlinear. In the past decades several algorithms have appeared to solve these equations. The first algorithm was SHAKE,[1] an iterative method for use with a leapfrog integrator. The equivalent for the velocity-Verlet integrator is called RATTLE.[2] Because of their iterative nature these algorithms do not lend them selves well for parallelization. In the simplest approach[3] communication is required at each iteration. For a molecule with all bonds constrained and a time step of 2 fs this leads to around 10 iterations and communication steps per MD time step, which is a much higher communication load than that of the other parts of the MD algorithm. In principle one could use the strategy for parallelization that will be described in this paper for the LINCS (linear constraint solver) algorithm[4] also for iterative methods. But the problem with that is that the number of iterations is not known a priori, and, therefore, the data that need to be communicated are not known when the domain decomposition is (re)made.

To avoid the issue of nonlinearity, the problem can be reduced to a linear matrix equation if the second derivatives of the constraint equations are set to zero. However, in a finite discretization scheme corrections are necessary to achieve accuracy and stability. Several methods have been proposed based on this linearization that have been termed promising for parallel simulations.[5−8] But none of these methods has been widely used, because the inherently unstable algorithms require some (periodically applied) corrections.

Since there are currently no practical parallel constraints algorithms, molecular simulation packages do not allow for constraints to cross node boundaries. For codes using domain decomposition this means that in practice only bonds involving hydrogens can be constrained, as such bonds only couple locally to one heavy atom. For particle or force decomposition codes, such as GROMACS 3.3,[9] it means that molecules with all bonds constrained cannot be split over processor boundaries. For a protein in water one can then not parallelize efficiently over more than a few processors.

Without constraints the fastest motions in molecular simulations are bond vibrations involving hydrogens, with a period of about 10 fs. When these bonds are constrained, the time step can be doubled, since the next fastest motion, bond vibrations involving only heavy atoms, have the fastest

* Corresponding author e-mail: hessb@mpip-mainz.mpg.de.

10.1021/ct700200b CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/01/2007

mode with a period of about 20 fs. How big the time step can actually be is difficult to determine. A nice, but deceptive property of Verlet type integrators, which are used by all major simulation packages, is that vibrations in harmonic potentials are integrated with exact energy conservation. Thus energy conservation is independent of the time step. However, the effective temperature at which the ensemble for harmonic modes is generated increases with the time step. For 20 and 10 time steps per oscillation the effective temperature is 3% and 14% too high, respectively. When such harmonic modes are not considered important, their temperature increase can be ignored, but in that case constraining them is better, since this avoids instabilities.

Additionally replacing bonds involving heavy atoms by constraints does not allow for an increase in time step, since also angle vibrations involving hydrogens have the shortest period of 20 fs. In the GROMACS package these vibrations can also be removed by replacing most hydrogen atoms by virtual interaction sites and constraining C−O−H angles.[10] Since then the fastest remaining modes have a period of about 45 fs, and this allows for an increase in a time step of slightly more than a factor of 2. Recently it was decided to implement domain decomposition into the GROMACS package, and, therefore, a parallel constraint algorithm is required, otherwise a factor of 2 in performance would be lost.

A decade ago the LINCS (linear constraint solver) algorithm was introduced.[4] This algorithm builds on the same linear approximation stated above but improves upon earlier algorithms in three ways. First a term is added to the equations which is analytically zero but for the discretized version leads to a completely stable algorithm. Second, iterations are applied to capture the nonlinear effects (i.e., bond rotations); under most circumstances a single iteration suffices. Third, the matrix is inverted efficiently using a series expansion, which leads to a bounded range of couplings between constraints, equal to the expansion order. The algorithm can be used for any type of integrator. In the original paper it is presented for the leapfrog integrator. But it is ideally suited for projecting out components of velocities or forces, a linear problem for which no iterations are required. The LINCS algorithm is the standard constraint algorithm in the GROMACS package, next to the SETTLE algorithm[11] which is only used for water molecules. Recently a “matrix-version’’ of the SHAKE algorithm has been developed,[12] which is parallelized with particle decomposition. This algorithm solves the same matrix equations as those of LINCS but uses a conjugate gradient solver. Therefore it does not have the exactly bounded coupling range that makes LINCS suitable for use in domain decomposition.

The original paper hinted that LINCS is easy to parallelize. That is what will be shown in the following, although in a slightly different way than originally stated. Additionally the energy conservation properties of LINCS are shown and compared with SHAKE in the light of an improvement for uncoupled angle constraints and recent improvements in integration accuracy.

II. The LINCS Algorithm

A concise description of the LINCS algorithm will now be presented, and the full derivation can be found in the LINCS paper.[4]

Consider a system of N particles, with positions given by a $3N$ vector $\mathbf{r}(t)$. The equations of motion are given by Newton’s law

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \mathbf{M}^{-1}\mathbf{f} \tag{1}$$

where $\mathbf{f}$ is the $3N$ force vector and $\mathbf{M}$ is a $3N \times 3N$ diagonal matrix, containing the masses of the particles. In general a system is constrained by K time-independent constraint equations

$$g_i(\mathbf{r}) = 0 \quad i = 1,..., K \tag{2}$$

The constrained system can still be described by $3N$ second-order differential equations in Cartesian coordinates.[13,14] The constraints will be applied according to the principal of least action.[15] In this approach the constraints are added as a zero term to the potential $\mathbf{V}(\mathbf{r})$, multiplied by Lagrange multipliers $\lambda_i(t)$

$$-\mathbf{M}\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \frac{\partial}{\partial \mathbf{r}}(\mathbf{V} - \lambda\cdot\mathbf{g}) \tag{3}$$

A new notation is introduced for the gradient matrix of the constraint equations which appears on the right-hand side of the equation

$$B_{hi} = \frac{\partial g_h}{\partial r_i} \tag{4}$$

Note that $\mathbf{B}$ is a $K \times 3N$ matrix, and it contains the directions of the constraints. Equation 3 can now be simplified to give

$$-\mathbf{M}\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} + \mathbf{B}^T\lambda + \mathbf{f} = 0 \tag{5}$$

The equations can be solved for λ to give the constrained equations of motion

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = (\mathbf{I} - \mathbf{TB})\mathbf{M}^{-1}\mathbf{f} - \mathbf{T}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} \tag{6}$$

where $\mathbf{T} = \mathbf{M}^{-1}\mathbf{B}^T(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T)^{-1}$. The projection matrix $\mathbf{I} - \mathbf{TB}$ projects out the components of a vector in the directions of the constraints, $\mathbf{M}^{-1}\mathbf{f}$ is the vector of unconstrained second derivatives, and $\mathbf{T}$ is a $3N \times K$ matrix that transforms motions in the constrained coordinates into motions in Cartesian coordinates, without changing the equations of motion of the unconstrained coordinates. The last term in (6) represents centripetal forces caused by rotating bonds. If the constraints are satisfied in the starting configuration, the linear differential eq 6 will conserve the constraints. The nonlinearity arises when eq 6 is discretized.

For holonomic constraints the constraint equations can be chosen as

$$g_i(\mathbf{r}_n) = |\mathbf{r}_{n,i_1} - \mathbf{r}_{n,i_2}| - d_i = 0 \quad i = 1,..., K \tag{7}$$

where d_i is the reference distance between atoms i_1 and i_2. The first step in the LINCS algorithm for a leapfrog integrator gives the linear, "zeroth iteration'' correction

$$\mathbf{r}^0_{n+1} = (\mathbf{I} - \mathbf{T}_n\mathbf{B}_n)\mathbf{r}^*_{n+1} + \mathbf{T}_n\mathbf{d} \tag{8}$$

where $\mathbf{r}^*_{n+1}$ is the unconstrained updated configuration:

$$\mathbf{r}^*_{n+1} = \mathbf{r}_n + \Delta t\mathbf{v}_{n-1/2} + (\Delta t)^2\mathbf{M}^{-1}\mathbf{f}_n \tag{9}$$

This algorithm is already stable but does not capture the nonlinear effect of the lengthening of constraints due to rotation. To correct for this, iterations are applied:

$$\mathbf{r}^{z+1}_{n+1} = (\mathbf{I} - \mathbf{T}_n\mathbf{B}_n)\mathbf{r}^z_{n+1} + \mathbf{T}_n\mathbf{p}^z \tag{10}$$

The projected lengths $\mathbf{p}^z$ are chosen by assuming that the observed displacements in iteration z perpendicular to the constraint direction in step n are (nearly) correct

$$p^z_i = \sqrt{2d_i^{\,2} - (l^z_i)^2} \tag{11}$$

where l^z_i is the slightly too long a distance in configuration $\mathbf{r}^z_{n+1}$. For most systems a single iteration provides sufficient accuracy. Quantitative results will be shown later. For projecting out constraint components of other quantities, such as velocities or forces, only the linear projection is required:

$$\mathbf{v}_n = (\mathbf{I} - \mathbf{T}_n\mathbf{B}_n)\mathbf{v}^*_n \tag{12}$$

The main computational issue is the matrix inversion required to obtain $\mathbf{T}$. This is simplified by left and right multiplying the constraint coupling matrix $\mathbf{B}_n\mathbf{M}^{-1}\mathbf{B}^T_n$ with a diagonal $K \times K$ matrix $\mathbf{S}$ containing the inverse square root of the diagonal of the coupling matrix:

$$\mathrm{S} = \mathrm{Diag}\left(\sqrt{\frac{1}{m_{1_1}} + \frac{1}{m_{1_2}}}, ..., \sqrt{\frac{1}{m_{K_1}} + \frac{1}{m_{K_2}}}\right) \tag{13}$$

The conversion goes as follows:

$$(\mathbf{B}_n\mathbf{M}^{-1}\mathbf{B}^T_n)^{-1} = \mathbf{S}\mathbf{S}^{-1}(\mathbf{B}_n\mathbf{M}^{-1}\mathbf{B}^T_n)^{-1}\mathbf{S}^{-1}\mathbf{S} = \mathbf{S}(\mathbf{S}\mathbf{B}_n\mathbf{M}^{-1}\mathbf{B}^T_n\mathbf{S})^{-1}\mathbf{S} \equiv \mathbf{S}(\mathbf{I} - \mathbf{A}_n)^{-1}\mathbf{S} \tag{14}$$

The matrix $\mathbf{A}_n$ is symmetric and sparse and has zeros on the diagonal. Thus a series expansion can be used to calculate the inverse:

$$(\mathbf{I} - \mathbf{A}_n)^{-1} = \mathbf{I} + \mathbf{A}_n + \mathbf{A}_n^{\,2} + \mathbf{A}_n^{\,3} + ... \tag{15}$$

The inversion only converges when the absolute values of all the eigenvalues of $\mathbf{A}_n$ are smaller than one. For calculating the expansion only matrix-vector multiplications are required. Since A is very sparse, only the nonzero elements should be stored, and the multiplications are computationally cheap.

Nearly all bonds in molecular simulations are sp^3 or sp^2 hybridized, which leads to a cosine of the angle between bonds of $-1/3$ and $-1/2$, respectively. The off-diagonal elements of A are given by this cosine times a mass factor which is between 0.5 and 1. This results in a maximum eigenvalue of A of 0.6−0.7, which means that the inversion always converges. The effective eigenvalue for typical bond distortions in MD simulations is around 0.4. Thus for each term in the expansion, the projection becomes more accurate by a factor of 0.4. The accuracy of each iteration can be less though, since it is limited by the guess for the projection of that iteration. The practical range for the expansion order is between 4 and 8. For large angle-constrained molecules eigenvalues larger than one occur, and therefore a different inversion method is required.

Nonconnected angle constraints appear in methyl, NH_3^+, and COH groups when angle vibrations involving hydrogens are removed.[10] Such individual angle constraints produce large eigenvalues with a localized eigenvector in matrix A. Especially the rigid COH group is problematic with a largest eigenvalue of 0.7, which is significantly larger than the effective eigenvalues of 0.4 for bond constraints. This imbalance means that the convergence of the expansion (15) can be limited by a few angle constraints. To avoid computational overhead due to angle constraints, the expansion can be extended for some couplings only

$$(\mathbf{I} - \mathbf{A}_n)^{-1} \approx \mathbf{I} + \mathbf{A}_n + ... + \mathbf{A}_n^{N_i} + (\mathbf{A}^*_n + ... + \mathbf{A}_n^{*N_i})\mathbf{A}_n^{N_i} \tag{16}$$

where N_i is the normal order of the expansion, $\mathbf{A}^*$ only contains the elements of $\mathbf{A}$ that couple constraints within rigid triangles, and all other elements are zero. In this manner the accuracy of angle constraints comes close to that of the other constraints, while the series of matrix vector multiplications required for determining the expansion only needs to be extended for a few constraint couplings.

The last point is how the constraints need to be applied to derivatives of the coordinates, namely the velocity and the forces. In GROMACS, as in most other simulations packages, the velocities were determined from the difference between the new and the old constrained positions. This can lead to inaccurate integration in single precision, since the increment of the coordinates can be very small. For the leapfrog integrator, eq 12 applied to the half step velocity would provide a more accurate solution. But recently it has been shown that one can also directly use the Lagrange multipliers.[16] For the LINCS algorithm these are already calculated, and the velocity correction then reads

$$\mathbf{v}_{n+1/2} = \mathbf{v}^*_{n+1/2} + \frac{1}{\Delta t}M^{-1}\mathbf{B}^T_n\lambda \tag{17}$$

Similarly the contribution of the constraints to the virial can be determined from the constraint forces which in the derivation above are given by the Lagrange multipliers. For the virial the inner product of the distance with the constraint forces $\mathbf{f}_c$ is required; this is simply $\mathbf{d}\cdot\lambda$. The full tensor can be obtained by using the outer products of the constraint directions with themselves.

III. The P-LINCS Algorithm

The inversion through a series expansion provides a nice physical picture. The coupling matrix A gives the direct coupling between bonds. The matrix A^2 gives the coupling between bonds separated by one bond and also the back-coupling of bonds to themselves. The matrix A^3 gives the

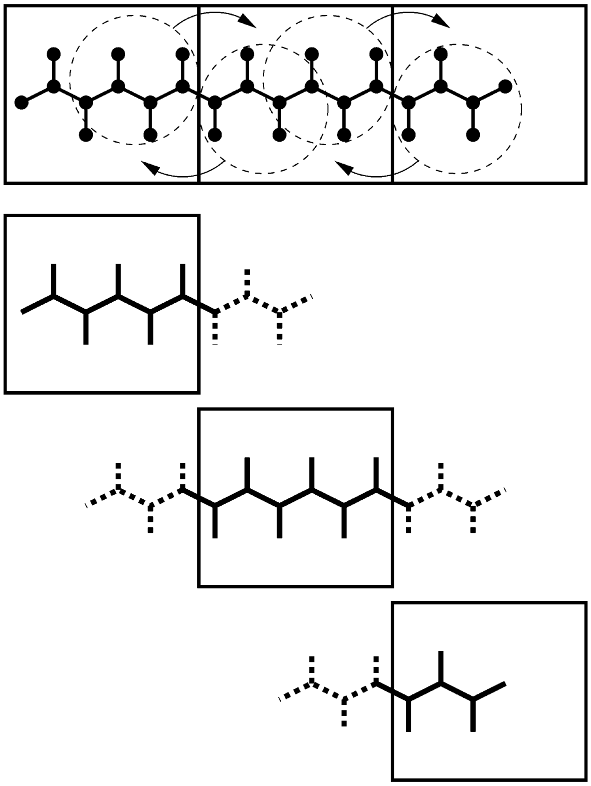


Figure 1. Example of the parallel setup of P-LINCS with one molecule split over three domain decomposition cells, using a matrix expansion order of 3. The top part shows which atom coordinates need to be communicated to which cells. The bottom parts show the local constraints (solid) and the nonlocal constraints (dashed) for each of the three cells.

coupling between bonds separated by two bonds, etc. This means that with this inversion method bonds do not influence each other when they are separated by more bonds than the maximum order in the expansion. This fact can be used to parallelize the algorithm.

When domain decomposition is used, each domain can apply the LINCS algorithm not only to the local constraints but also to some constraints involving only atoms of the neighboring domains. In the original paper it was suggested to communicate the atoms for the extra constraints for all iterations at once. But it turns out to be more practical, and often also more efficient, to communicate before each iteration. Then in addition to the atoms of constraints that cross the borders between domains, atoms that are separated by maximally the number of bonds equal to the order of the expansion need to be communicated (see Figure 1). When these coordinates are present, a LINCS iteration can be applied to local plus extra constraints. After such an iteration, the local constraints and the ones crossing the border will be in an identical state to a nonparallel version of the algorithm. The extra, nonlocal constraints will differ, since they have not felt the influence of some coupled constraints, but this does not matter, since in the end state they do not influence the local atoms. Note that for determining the constraint contribution to the virial, one should take care that constraints that cross cell boundaries are not double counted. The additional terms in the expansion (16) for angle constraints can interfere with the exact correspondence of P-LINCS and LINCS results but only when two or more triangles of constraints have a constraint in common. However, this is not an issue, since for such cases the matrix expansion converges too slowly or not at all. The procedure for removing angle vibrations of hydrogens[10] does not introduce coupled triangles of constraints.

Before each iteration the updated nonlocal coordinates need to be communicated. This leads to a total number of communication steps of one plus the number of iterations. No communication is required after the last iteration. Since the number of iterations is usually one, the number of communication steps is two. By doing a communication step for each iteration an expansion order of up to 6 can be used, since the number of bonds in an all-trans chain that fits in a domain decomposition cell size of 1 nm (a typical minimum value) is 7. When required, extra communication steps can be added for atoms in more distant cells. The same procedure can be used for constraining velocities or forces; there only one communication step is required.

The final result of the P-LINCS algorithm is identical to that of the LINCS algorithm, save for numerical rounding differences. In the implementation of P-LINCS in the upcoming 4.0 version of the GROMACS package, the coupling matrix is always stored in the same order, which leads to binary identical results to those of LINCS.

When the required extra pieces of molecule(s) are not longer than the smallest dimension of a domain decomposition cell, one cell needs to communicate with at most 26 other cells with full 3D domain decomposition. The communication can be performed in 3 steps of pairs of communication calls. In the P-LINCS implementation in GROMACS the constraint communication setup is redetermined every time the decomposition changes, which is usually every 5−10 integration steps. Every cell has a list of all the constraints in the whole system. Each cell can then determine of which nonlocal atoms it needs the coordinates for connected constraints that cross the cell boundaries. The list of required atoms is first sent one cell forward and backward in the *x* direction. Atoms in the received list that are locally present are marked, and the rest, in addition to the locally required atoms, are sent forward and backward one cell in the *y* direction. The same procedure is repeated for the *z* direction. In the opposite order and direction the cells send and accumulate the found atom indices. Each cell can then determine if all required atoms have been found. Before each LINCS iteration the last part of the procedure is repeated but then with the coordinates instead of the atom indices. This results in a maximum of 6 communication steps per iteration. For a machine with two-way network connections the forward and backward calls can be overlapped. The required bandwidth will be quite low compared to the latency. The passing of coordinates through other cells leads to little overhead, since these are usually small in number, and often these coordinates are also required by the cells they pass through.

IV. Benchmarks

It is difficult to assess the accuracy of MD simulations of biomolecular systems. Ideally one would want to check how accurate the generated ensemble is, but for proteins and even for peptides of more than a few amino acids it is computationally infeasible to sample the full phase space. A constraint algorithm should, of course, accurately set the constraint lengths. But how accurate is accurate enough, a relative error of 10^{-4}, or maybe 10^{-6}? An easy quantitative check is the accuracy of energy conservation. But also for this quantity is it difficult to judge what value is required. For microcanonical simulations one can easily determine how much drift in the total energy one can allow over the total simulation time. But most simulations are performed in the canonical or constant-NPT ensembles. Here the thermostat will effectively compensate for energy changes due to (small) integration errors. How much energy drift can be allowed is therefore unclear. Also one should keep in mind that Verlet type integrators perfectly conserve energy for harmonic potentials for any and therefore also unrealistically large, integration step size. This means that for the accuracy of especially simulations without bond constraints one cannot rely on energy conservation as a measure of integration accuracy.

It is clear that a good algorithm should be able to reach any energy conservation value required by the user. To demonstrate this for LINCS, we simulated the actin-binding domain of villin headpiece (36 residues) with the OPLS all-atom force field.[17] To avoid cutoff artifacts all simulations were performed in vacuo without cutoffs. To ensure the same conditions for all LINCS accuracies, canonical simulations at 300 K were performed using a Nose-Hoover thermostat, implemented in GROMACS with a reversible leapfrog integrator.[18] The period of the temperature oscillations was set to 2 ps. The energy conservation accuracy is obtained from the drift of the conserved energy quantity in Nose-Hoover dynamics. The results for LINCS and SHAKE for simulations of 1 ns are shown in Table 1. One can see that with a logarithmic increase in computational effort the accuracy of LINCS can be increased to a finally unmeasurable drift over 1 ns in double precision. The amount of drift can be compared to another common source of drift, namely cutoff artifacts. Often plain cutoffs are use for nonbonded interactions, mainly for reasons of computational efficiency. A plain cutoff for the Lennard-Jones interactions of 0.9−1.1 nm with a neighbor-list update interval between 10 and 20 fs introduces energy into the system at a rate of of 1−10 k_BT/ns per degree of freedom. Reaction-field electrostatics produces 1 order of magnitude more drift.

The amount of time spent in the constraint algorithm is relatively high in this system. For a protein in solvent the relative time for LINCS will be a factor of 2−4 lower. In single precision it does not make sense to increase the order of matrix expansion above 6, since all further terms in the matrix are beyond the numerical precision compared to the first terms. Already for the case with two iterations and expansion order 6 the energy drift is unmeasurable in single precision, whereas in double precision there is a clear, although very small, negative drift. This is because in single precision numerical rounding errors cancel the small analytical error of the LINCS algorithm. The effect of the old way of constraining the velocities, using the changes in coordinates, can only be observed for these same settings. LINCS is computationally slightly more efficient than SHAKE, but this difference can probably be overcome by using overrelaxation.[19] When the original LINCS paper was written, a decade ago, LINCS was a factor of 4 faster than SHAKE. This difference has become much smaller, because modern processors can more efficiently execute code with conditional statements such as SHAKE.

A clear advantage of LINCS over SHAKE is its stability at large time steps. To show the performance of LINCS with large time steps villin was simulated with all hydrogens replaced by virtual sites or angle constraints[10] with time steps of 2, 4, and 5 fs. The results are shown in Table 2. For maintaining the constraint and integration accuracy with increasing time step, the order of the expansion and/or the number of LINCS iterations need to be increased. This also increases the computational effort slightly, but LINCS still takes a negligible amount of the total run time.

To illustrate the performance of P-LINCS, T4-lysozyme in a rectangular box of 5.2 × 6.7 × 5.2 nm ^{3} with 5000 SPC water molecules[20] and 8 Cl^- ions was simulated with

Table 1. Accuracy of the LINCS Algorithm Applied to Villin with a Time Step of 2 fs in Single and Double Precision as a Function of the Number of Iterations N_i and the Order of the Expansion, in Terms of the Relative Root-Mean-Square Deviation (RMSD) of the Constraint Lengths and the Drift of the Conserved Energy in k_BT per Degree of Freedom[a]

						energy		
				tol.,	RMSD,	drift,	time,	time,
	prec.	N_i	order	10^{-4}	10^{-4}	ns^{-1}	ms	%
LINCS	single	1	4		0.27	−1.59	0.14	3.9
	single	1	6		0.11	−0.34	0.17	4.8
	single	2	6		0.02	0.00	0.24	6.5
	double	2	6		0.012	−0.03	0.30	4.8
	double	2	8		0.003	0.00	0.35	5.6
LINCS	single	1	4		0.27	−1.60	0.15	3.9
old	single	1	6		0.11	−0.35	0.17	4.8
v corr.	single	2	6		0.02	−0.08	0.24	6.5
SHAKE	single			1.00		−2.06	0.15	4.3
	single			0.10		−0.06	0.22	6.1
	double			0.10		−0.14	0.28	4.6
	double			0.01		−0.01	0.39	6.2

[a] The last column shows the CPU time spent in the constraint algorithm per step and as a percentage of the total run time. For comparison LINCS with the old, inaccurate velocity correction and SHAKE with difference relative constraint tolerances are also shown. Benchmarks were performed on one core of an Intel 2.4 GHz Core 2 CPU.

Table 2. As Table 1, but Only for LINCS in Single Precision for Villin with Virtual Interaction Sites

			RMSD,	energy drift,	time,	time,
Δt	N_i	order	10^{-4}	ns^{-1}	ms	%
2	1	4	0.22	−1.26	0.10	2.5
2	1	6	0.08	−0.23	0.13	3.7
4	1	6	0.31	−1.67	0.13	3.7
4	2	6	0.03	−0.15	0.18	5.1
5	1	6	0.50	−3.07	0.13	3.7
5	2	6	0.05	−0.26	0.18	5.1

Table 3. Performance of P-LINCs on Lysozyme in Water (See Text for Details) without and with Virtual Sites as a Function of the Time Step (Δt) and the Number of Domain Decomposition Cells, except for pd4 Which Is Particle Decomposition over 4 Processors[a]

					constraint time (ms)					speed (ns/day)				
virtual sites	Δt, fs	N_i	order	RMSD, 10^{-4}	1	pd4	4	16	32	1	pd4	4	16	32
no	2	1	4	0.21	2.4	2.4	3.8	5.7	8.8	3.0	10.2	11.8	41	70
yes	2	1	4	0.26	2.3	2.3	3.3	4.9	7.6	3.0	10.2	11.8	40	69
yes	2	1	6	0.07	2.5	2.5	3.6	5.9	9.0	3.0	10.2	11.6	40	67
yes	4	1	6	0.29	2.5	2.5	3.6	5.9	9.0	5.3	18.1	21.2	70	117
yes	5	1	6	0.47	2.5	2.5	3.6	5.9	9.0	6.3	21.5	25.1	83	137

[a] RMSD is the relative deviation of the constraint lengths, constraint time gives the time used by the constraint algorithms (LINCS/P-LINCS plus settle for water) in one integration step summed over all the processors, and speed is the simulation time per day. Benchmarks were performed on a 2.2 GHz AMD64 cluster with Infiniband interconnects.

the GROMOS 53a6 force-field[21] with united aliphatic carbons. To create a computationally demanding test case for P-LINCS, a plain cutoff of 1.1 nm was used for the Lennard-Jones interaction and the reaction-field electrostatics, with a neighbor-list update every 20 fs. These settings do not lead to very good energy conservation. Tapered cutoffs and/or particle-mesh Ewald electrostatics provide much better energy conservation, but are, in the GROMACS package, computationally roughly twice as expensive, since tabulated instead of analytical potentials need to be used. This will lower the relative computational cost of P-LINCS and is therefore a less demanding benchmark for P-LINCS.

Timings for P-LINCS only and the complete simulation, without and with virtual sites, are shown in Table 3 using a preliminary version of the GROMACS 4.0 package with load-balanced domain decomposition. The order of the expansion for P-LINCS needs to be adjusted with increasing time step, which leads to a marginally higher computational cost. The time for the constraint algorithm includes the time used by SETTLE for the water molecules. This combination of algorithms does not cause load imbalance, since SETTLE does not communicate and is done after P-LINCS. Processors that have a lot of water molecules to constrain have less protein constraints, and the two algorithms roughly balance out during the last LINCS iteration. One can see that the time for P-LINCS increases with an increasing number of processors/cells. This is not due to the extra constraints across the cell boundaries but nearly only due to communication. When going from 4 to 32 processors the domain decomposition goes from 1D to 3D, requiring from 1 to 3 communication steps per iteration. Since the time for a communication step in P-LINCS, which is latency bound, stays nearly constant, the P-LINCS time increases with the dimensionality of the domain decomposition. However, even on 32 processors P-LINCS still only takes 10% of the computation time. This number will halve when more accurate cutoff schemes are employed, even when an extra P-LINCS iteration is used for more accuracy. For comparison LINCS results are shown with particle decomposition on 4 processors, which is slightly slower than domain decomposition. Particle decomposition without parallel constraints is limited to 4 processors, since already on 5 processors there is a load imbalance when one processor needs to take care of the whole protein. It should also be noted that with a time step of 5 fs the neighbor list is updated every 4 steps, which is costly. In practice one would increase the neighbor-list cutoff and decrease the neighbor-list update frequency to improve perform ace. This was not done here to keep the simulations comparable.

As mentioned before, the P-LINCS communication is latency limited. For the lysozyme system with a matrix expansion order of 6, the number of atoms to be communicated between the different cell boundaries varies between 0 and 330 for 4 cells and between 0 and 130 for 16 or 32 cells. This means that in single precision the largest message is 4000 bytes for 4 cells and 1600 bytes for 16 or 32 cells. With an all-atom force field there are twice as many constraints, and, therefore, the numbers double. With hydrogen angle vibrations removed the numbers are twice as small, also for an all-atom force-field, since most constraints involving hydrogens dissappear. For such message sizes the communication is latency limited on any system. The message size depends mainly on the cell size. As the cell size increases, the message size also increases. Thus for large cells the communication may no longer be latency limited. But since the computational cost for the forces and constraints is proportional to the volume of the cell and the number of atoms for constraint communication proportional to its surface, the P-LINCS communication will never be a limiting factor.

V. Conclusions

P-LINCS is a parallel constraint algorithm which gives (binary) identical results to the nonparallel algorithm LINCS. It therefore has the same high stability. Its implementation on top of LINCS is straightforward, requiring only the coding of some bookkeeping and communication. The treatment of uncoupled angle constraints has been improved compared to the original version of the LINCS algorithm. The computational cost of P-LINCS increases with the dimensionality of the domain decomposition grid, since the communication is latency limited. But the time spent in P-LINCS is negligible on the total run time. The computational cost increases logarithmically with the required accuracy. For a typical protein simulation with the GROMACS package the total cost of P-LINCS with full 3D domain decomposition is between 4% and 10%, depending on the treatment of the electrostatics. In other packages this could be significantly less, since in GROMACS the force calculation is very efficient due to assembly SSE/SSE2 force loops.

Acknowledgment. The author thanks David van der Spoel and Erik Lindahl for stimulating discussions.

References

(1) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(2) Andersen, H. *J. Comput. Phys.* **1983**, *52*, 24.

(3) Brown, D.; Clarke, J. H. R.; Okuda, M.; Yamazaki, T. *Comput. Phys. Comm.* **1994**, *83* (1), 1.

(4) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.

(5) Edberg, R.; Evans, D. J.; Morriss, G. P. *J. Chem. Phys.* **1986**, *84*, 6933.

(6) Baranyai, A.; Evans, D. J. *Mol. Phys.* **1990**, *70*, 53.

(7) Yoneya, M.; Berendsen, H. J. C.; Hirasawa, K. *Mol. Simul.* **1994**, *13*, 395.

(8) Slusher, J. T.; Cummings, P. T. *Mol. Sim.* **1996**, *18*, 213.

(9) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.

(10) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.

(11) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952.

(12) Weinbach, Y.; Elber, R. *J. Comput. Phys.* **2005**, *209*, 193.

(13) de Leeuw, S. W.; Perram, J. W.; Petersen, H. G. *J. Stat. Phys.* **1990**, *61*, 1203.

(14) Bekker, H. Molecular Dynamics Simulation Methods Revised, Ph.D. Thesis, University of Groningen, The Netherlands, 1996.

(15) Landau, L.; Lifshitz, E. *Mechanics;* Pergamon Press: Oxford, 1961.

(16) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Shaw, D. E. *J. Chem. Phys.* **2007**, *126*, 046101.

(17) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(18) Holian, B. L.; Voter, A. F.; Ravelo, R. *Phys. Rev. E* **1995**, *52* (3), 2338.

(19) Barth, E.; Kuczera, K.; Leimkuhler, B.; Skeel, R. D. *J. Comput. Chem.* **1996**, *16*, 1192.

(20) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, 1981; pp 331−342.

(21) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656.

CT700200B

JCTC Journal of Chemical Theory and Computation

TD-DFT Performance for the Visible Absorption Spectra of Organic Dyes: Conventional versus Long-Range Hybrids

Denis Jacquemin,*[,†] Eric A. Perpète,[†] Gustavo E. Scuseria,[‡] Ilaria Ciofini,[§] and Carlo Adamo*[,§]

Laboratoire de Chimie Théorique Appliquée, Groupe de Chimie-Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium, Department of Chemistry, Rice University, Houston, Texas 77005, and Ecole Nationale Supérieure de Chimie de Paris, Laboratoire Electrochimie et Chimie Analytique, UMR CNRS-ENSCP no. 7575, 11, rue Pierre et Marie Curie, F-75321 Paris Cedex 05, France

Received July 28, 2007

Abstract: The $\pi \rightarrow \pi^*$ transitions of more than 100 organic dyes from the major classes of chromophores (quinones, diazo, ...) have been investigated using a Time-Dependent Density Functional Theory (TD-DFT) procedure relying on large atomic basis sets and the systematic modeling of solvent effects. These calculations have been performed with pure (PBE) as well as conventional (PBE0) and long-range (LR) corrected hybrid functionals (LC-PBE, LC-ωPBE, and CAM-B3LYP). The computed wavelengths are systematically guided by the percentage of exact exchange included at intermediate interelectronic distance, i.e., the $\lambda_{\max}$ value always follows the PBE > PBE0 > CAM-B3LYP > LC-PBE > LC-ωPBE > HF sequence. The functional giving the best estimates of the experimental transition energies may vary, but PBE0 and CAM-B3LYP tend to outperform all other approaches. The latter functional is shown to be especially adequate to treat molecules with delocalized excited states. The mean absolute error provided by PBE0 is 22 nm (0.14 eV) with no deviation exceeding 100 nm (0.50 eV): PBE0 is able to deliver reasonable estimates of the color of most organic dyes of practical or industrial interest. By using a calibration curve, we found that the LR functionals systematically allow an even more consistent description of the low-lying excited-state energies than the conventional hybrids. Indeed, linearly corrected LR approaches yield an average error of 10 nm for each dye family. Therefore, when such statistical treatments can be designed for given sets of dyes, a simple and rapid theoretical procedure allows both a chemically sound and a numerically accurate description of the absorption wavelengths.

I. Introduction

Though dyes could be classified with respect to the chemical process generating the color (absorption/fluorescence/phosphorescence) or to the nature of the implied excited states ($\pi \rightarrow \pi^*/n \rightarrow \pi^*$), one generally groups them according to the nature of their chromophoric unit (Figure 1).[1−3] The two major families of organic dyes with industrial applications are 9,10-anthraquinones (AQ) and azobenzenes (AB), that represents about 30% and 60% of today's world dye production, respectively.[1−3] The longest wavelength of maximal absorption ($\lambda_{\max}$) of AQ covers all the visible region of the electromagnetic spectrum, depending on the nature

* Corresponding author e-mail: denis.jacquemin@fundp.ac.be, http://perso.fundp.ac.be/~jacquemd (D.J.), carlo-adamo@enscp.fr (C.A.).

[†] Facultés Universitaires Notre-Dame de la Paix.

[‡] Rice University.

[§] UMR CNRS-ENSCP no. 7575.

10.1021/ct700187z CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/17/2007

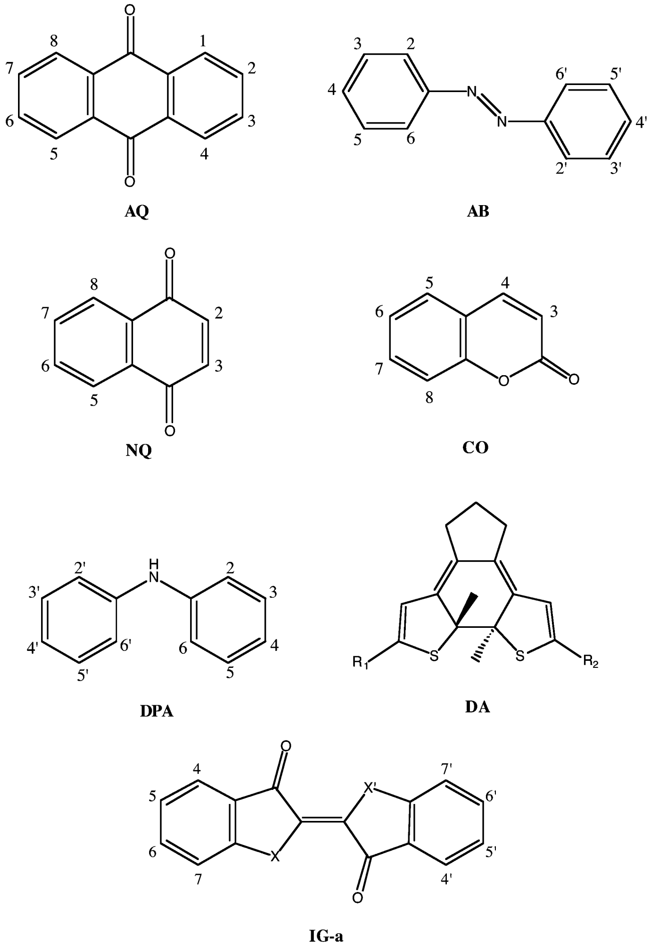


Figure 1. Sketch of the chromophores investigated in this study.

and position of the auxochromic group(s) substituting positions 1-to-8.[2,4] AB are extremely versatile,[3] with applications going from core of media storages,[5] to central building blocks in molecular motors.[6,7] Of course, several other chromophores related to more specific applications can be pinpointed: (1) naphthoquinones (NQ), implied in several medicinal processes;[8,9] (2) coumarins (CO), the most efficient fluorescent brighteners;[10] (3) diphenylamine derivatives (DPA), the typical hair dyes with important biological properties;[11,12] (4) diarylethenes (DA), the prototype molecular switch;[13−15] and (5) indigoids derivatives (IG, Figure 2) which give loads of structures with several substitution patterns of the outer phenyl rings, of the heteroatoms, as well as different types of linkage between the two parts of the molecule.[3] Developing molecular modelization approaches allowing an accurate prediction of the color of dyes is still a major challenge,[16] because, on the one hand, the average human eye is able to tell apart shades differing by 1 nm only, and, on the other hand, actual stains are medium-sized molecules, possess a dozen π-electrons, and are very sensitive to the environments. Therefore, large-scale highly correlated ab initio approaches such as EOM-CC, MR-CI, or CAS-PT2 remain out of today's computational reach. Consequently, one could be inclined to select customized semiempirical approaches such as ZINDO, but the consistency of such schemes is often disappointing.[17−19] Currently, the most widely applied ab initio tool for modeling electronic spectra of structures is the time-dependent density functional theory (TD-DFT).[20] TD-DFT calculations can incorporate environmental effects[21] and quickly give UV/vis spectra for most organic[22−25] and inorganic[26,27] dyes. Still, meaningful results can only be attained with a selection of adequate exchange-correlation functionals. It is recognized that conclusions obtained with hybrid functionals tend to be in better agreement with experimental trends than the values computed with pure functionals. Hybrids, originally proposed in the 1990s,[28,29] include a fraction (α) of *exact* exchange that is computed with the Hartree−Fock (HF) exchange formula.[28−40] Despite their countless successes, hybrids also encounter problems that seem (mostly) independent of the functional selected. Typical troublesome properties include van der Waals forces,[41] bond length alternation (BLA) in semiconducting polymers,[42,43] nonlinear optics (NLO) properties of long π-conjugated chains,[44,45] and charge-transfer electronic transitions.[17,46−49] In these four cases, no single α value provides a small (or consistent) error for increasingly large/spaced compounds. In fact, these DFT limitations have a common origin: the so-called shortsightedness of DFT functionals. In other words, the density is not influenced by a change in the nearby electronic distribution.[44,47,48] To circumvent these shortcomings, several strategies have been designed and applied to the problems listed above: the correction(s) of the self-interaction error,[50−52] the inclusion of the current-density in the formalism,[53,54] the addition of empirical dispersion terms,[55,56] and the use of optimized effective potential for exact exchange[57,58] as well as the explicit consideration of long-range effects (LR).[59−86] This latter scheme leads to the *range-separated* hybrids that use a growing fraction of exact exchange when the interelectronic

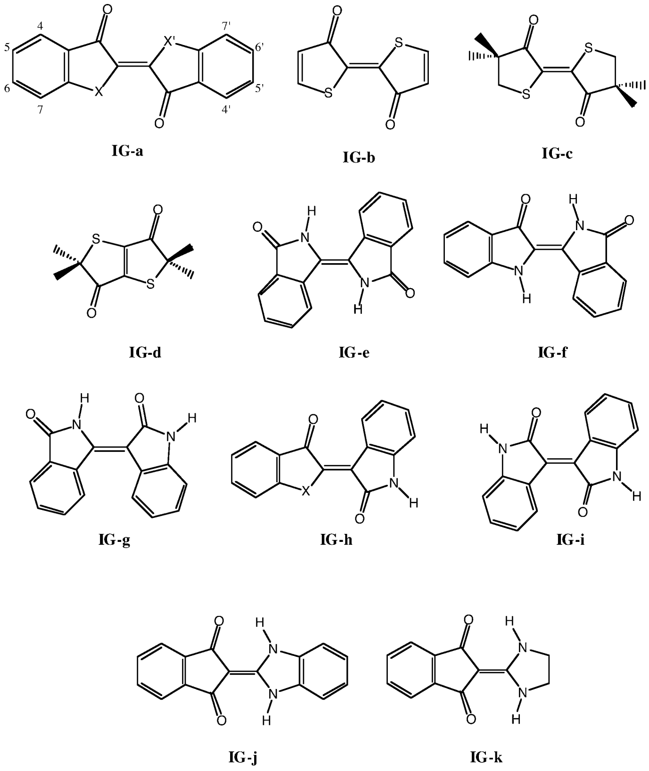


Figure 2. Studied indigoids derivatives.

distance increases (see section II). In contrast, the hybrids in which the amount of HF exchange is constant all over the space will be referred to as *global* hybrids in the following (*conventional* or *full-range exchange* hybrids have also been used in the literature). It has been demonstrated that range-separated hybrids are very efficient for calculating BLA[82] or NLO[61,67,68,82,83,87] properties in conjugated polymers as well as for determining properties of weakly bond complexes[64,70,85] or charge-transfer states in large molecular systems.[64,68,74−77,81] Nevertheless, there have been only a few works establishing the abilities of TD-LR-DFT to reproduce experimental UV/vis spectra for a statistically meaningful set of compounds: (1) comparisons of global and range-separated functionals performances for the vertical transitions of C═O, N_2, C_2H_4, H_2O, C_6H_6, and H_2C═O demonstrated that, while Rydberg's states are much better described with the latter, differences remain small for valence excited states;[64,73,85] (2) the emission properties of low-lying excited-states of small molecules have also been investigated, and similar conclusions have been drawn;[74] (3) we have studied the localized $n \rightarrow \pi^*$ transitions in nitroso and thiocarbonyl dyes, and it turned out that all hybrid functionals lead to a quite similar accuracy;[88] (4) the $\lambda_{\max}$ of four CO dyes are more accurate with global TD-DFT than with (unmodified) TD-LR-DFT;[84] (5) for the $\pi \rightarrow \pi^*$ transitions of 15 AQ, it has been found that range-separated hybrids are further away from experimental values than their global counterparts but offer a much smaller statistical dispersion of the results, allowing more valid chemical insights;[82] and (6) on the contrary, TD-LR-DFT brings no significant correction for cyanine derivatives, as these dyes present a strong multideterminantal nature.[82]

In this paper, we perform a critical assessment of the efficiency and consistency of range-separated hybrids for computing the main $\pi \rightarrow \pi^*$ transition of industrial organic dyes. The generic chromophores we have considered are depicted in Figure 1. It is worth pointing out previous TD-DFT investigations for these compounds. For AQ, the performance of global hybrids has been assessed in refs 19, 22, 46, and 89−93, while our previous work used TD-LR-DFT.[82] Numerous computations of the UV/Visible spectra of AB based on TD-DFT have been published,[46,94−98] but to our best knowledge none used range-separated hybrids. NQ, DPA, and IG have recently been tackled by two of us in refs 99, 100, 101, and 102−105, respectively. For CO, one finds several investigations performed with global hybrids,[106−113] but only one used range-separated functionals and was limited to four molecules.[84] The transition spectra of the photochromic DA switches have also been thoroughly investigated, though only with global hybrids.[15,18,114−123] In fact, the molecules in Figures 1 and 2 include most of the families selected by Guillaumont and Nakamura[46] (we excluded cyanine-like dyes that present a multideterminental nature) but with (much) more structures in each subset.

This paper is organized as follows. In section II, we briefly summarize our computational approach. In section IIIA, the spectra of the various families computed with several pure, global, and range-separated hybrids are compared to experimental data. In section IIIB, we examine the possibilities of statistical treatment of the theoretical $\lambda_{\max}$, before concluding in section IV.

II. Methodology

In range-separated functionals, the Coulomb operator is partitioned as[59,60,63,65]

$$\frac{1}{r_{12}} = \frac{1 - [\alpha + \beta \operatorname{erf}(\omega r_{12})]}{r_{12}} + \frac{\alpha + \beta \operatorname{erf}(\omega r_{12})}{r_{12}} \qquad (1)$$

where ω is the range separation parameter, while α and α + β define the exact exchange percentage at $r_{12} = 0$ and $r_{12} = \infty$, respectively. In eq 1, $0 \le \alpha + \beta \le 1$, $0 \le \alpha \le 1$, and $0 \le \beta \le 1$, are three conditions to be satisfied. Equation 1 leads to the partitioning of the total exchange energy into short-range and long-range contributions:

$$E_x = E_x^{\mathrm{sr}} + E_x^{\mathrm{lr}} \qquad (2)$$

In this paper, three range-separated functionals have been used: (1) the LC (LC: long-range correction) scheme of Hirao[61] applied to the PBE functional,[124] (2) the LC-ωPBE functional by Vydrov and Scuseria,[77] and (3) Handy's CAM-B3LYP (CAM-B3LYP: Coulomb-attenuating method applied to B3LYP).[65] Both LC models use $\alpha = 0$ and $\beta = 1$ in eq 1, i.e. short-range semilocal DFT exchange is combined with long-range HF exchange integrals. Since $\alpha + \beta = 1$, the exchange potential in LC functionals has the exact asymptotic behavior. Note that in LC-ωPBE, the short-range exchange functional can be rigorously derived[62,125] by integration of the model exchange hole.[77,78] In CAM-B3LYP, $\alpha = 0.19$ and $\beta = 0.46$ are plugged in, and the exact asymptote of the exchange potential is lost, while a larger percentage of HF exchange is included at short range. The range separation parameter, ω in LC-PBE and CAM-B3LYP, is set to the standard 0.33 bohr^{-1} value, whereas for LC-ωPBE, we use the optimized 0.40 bohr^{-1} value from refs 77 and 78. Recently, such 0.40 bohr^{-1} value has been advocated by Fromager and co-workers,[86] whereas Hirao et al. proposed a reoptimized value of 0.47 bohr^{-1} for reaction barriers.[126] As our goal is to assess the merits of range-separated and global hybrids for visible spectra simulations, we have also performed time-dependent calculations with a pure functional (PBE),[124] a global hybrid (PBE0, that contains 25% of exact exchange),[33,34] and the HF approach (in this paper HF results are obtained through the TD-HF approach). Readers interested in the results of other global hybrids, such as the archetype B3LYP, are refered to 89 for AQ, 109 for CO, and 105 for IG. The evolution with r_{12} of the exact exchange percentage used in the six considered models is sketched in Figure 3All calculations have been performed with the Gaussian03 suite of programs,[127] except for the LR-DFT calculations that were carried out with a development version of Gaussian,[128] using their standard TD-DFT procedure (ref 129). For each system, the ground-state structure has been determined by a standard force-minimization process, and the vibrational spectrum has been determined to systematically check that all vibrational frequencies are real. All these ground-state calculations have been performed with PBE0 using a triple-ζ polarized basis set, 6-311G(d,p), that

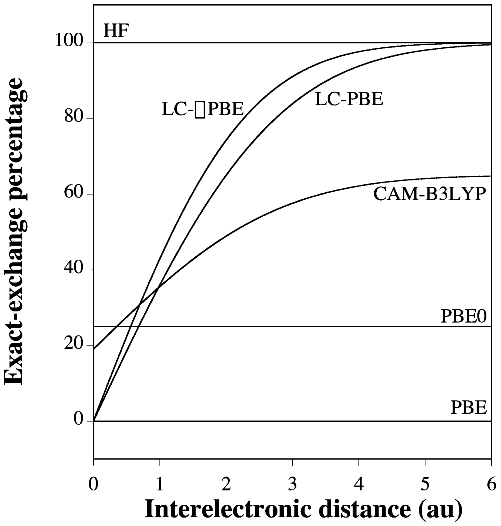


Figure 3. Evolution of the percentage of exact exchange included as the function of the interelectronic distance for the six models considered.

is known for providing converged ground-state structural parameters for the largest majority of the compounds.[130−132] In previous investigations, we have demonstrated that PBE0/6-311G(d,p) geometries are perfectly adequate for most classes of organic dyes investigated here,[18,99,101−105,122] and we refer the interested readers to these publications for discussion of the basis set effects. TD-DFT is then used to compute the three-to-eight first low-lying excited states of each dye. The resulting electronic excitations have a strong $\pi \rightarrow \pi^*$ character associated with a large oscillator force. We have systematically selected the 6-311+G(2d,p) basis set for these TD-DFT calculations, as it yields perfectly converged λ_{max} for IG,[102,103] DPA,[101] NQ,[99] and DA[18,122] dyes. For AQ, a smaller basis set would even be enough to attain the saturation of transition energies.[89,90] Therefore, we are very confident that all results presented here would not have been significantly affected by a further extension of the basis sets. At each step, the surrounding effects have been included by means of the Polarizable Continuum Model,[133] as valuable theory/experiment comparisons indeed require simulation of the solvent. Two models have been used the default IEF-PCM and the conducting PCM model (C-PCM). Computational details might slightly vary from one dye family to another (radii used to build the cavity, use of smoothing spheres, etc.) because we have set these computational parameters in order to maintain consistency with previously published data. Nevertheless, this should have a completely negligible impact on the computed wavelengths.[101] In this paper, we have selected the so-called nonequilibrium procedure for TD-DFT calculations, that has been specifically designed for the study of absorption processes.[21]

In many cases, several experimental values are available, and the values reported in Tables 1−5 correspond to the average measure. The selection of the theoretical wavelength is often straightforward: it is the first transition with a significant oscillator force.[19,99−105,110,123] In fact, to perfectly simulate experimental results, the main missing components are the vibronic couplings. Indeed, in some cases, the inclusion of Franck−Condon factors could be essential to get the best theory/experiment match.[113] However, a systematic computation of such vibronic effects is not practically feasible for our very extended set.

III. Results

A. Comparisons with Experiments. The λ_{max} computed for 24 typical AQ dyes are reported in Table 1. For all compounds, the absorption wavelengths systematically obey: PBE > PBE0 > CAM-B3LYP > LC-PBE > LC-ωPBE > HF. This means that the larger the exact exchange ratio at intermediate r_{12} (see Figure 3), the smaller the calculated λ_{max}. For nitroso and thiocarbonyl compounds, such a systematic relationship could not be unravelled,[88] probably due to the more localized nature of the transition in these $n \rightarrow \pi^*$ chromophores: the mixing percentage at smaller distances had a larger influence. Consistently with our previous studies,[19,82,89,90] PBE0 yields λ_{max} in very good agreement with experimental trends for the short-wavelength dyes, but the discrepancies significantly increase for the compounds with the smallest transition energies. Indeed, in the lower part of Table 1, it is PBE that yields the best estimates. For the 24 AQ, we obtain mean signed errors (MSE, experiment-theory) of 127, −71, 12, 67, 85, and 53 nm for HF, PBE, PBE0, LC-PBE, LC-ωPBE, and CAM-B3LYP, respectively. The corresponding mean absolute errors (MAE) amount to 127, 74, 19, 67, 85, and 53 nm, indicating that LR-DFT and HF systematically underestimate the λ_{max}. Consistently with the findings of ref 82, PBE0 is clearly closer to experiment, with a MAE less than half of the second competitor, namely CAM-B3LYP. However, the ordering of the compounds is also crucial for an efficient molecular design. Range-separated functionals provide the valid 1,4-OH > 1-NH_2 classification whereas PBE0 does not, but the reverse situation also appears (1,2-OH versus 1,8-OH), and cases can also be noted in Table 1 for which all approaches fail (2-OMe versus 1,2-OMe).

The situation differs in Table 2 where the spectra of AB derivatives are listed. Model and real-life AB have been considered, though we have not included OH substituents in the panel as such hydroxy-AB tend to undergo tautomerism that might impede straightforward theory/experiment comparisons. On the contrary, several push−pull molecules (4-NO_2, 4′-NH_2, and alike) having a strong charge-transfer character are tackled in Table 2. As for AQ, the methodological ordering of λ_{max} follows the amount of exact exchange at medium interelectronic distance. Nevertheless, CAM-B3LYP has now a slight edge over PBE0, and the accuracy difference between the λ_{max} obtained with global and range-separated approaches becomes less striking than for the AQ listed in Table 1. Indeed, we obtain MAE (MSE) of 64 (64) nm, 90 (−90) nm, 25 (−20) nm, 33 (33) nm, 46 (46) nm, and 20 (15) nm for HF, PBE, PBE0, LC-PBE, LC-ωPBE, and CAM-B3LYP, respectively. In fact CAM-B3LYP is particularly efficient for structures presenting a small λ_{max}, and the theory/experiment discrepancy tends to increase when going down the column. Note that for both AB having a λ_{max} > 500 nm, PBE0 is closer to the experimental reference than the three range-separated functionals: the description of charge-transfer dyes is not

Table 1. 9,10-Anthraquinone (AQ) Main Visible Transition, Obtained with the IEF-PCM-TD-X/6-311+G(2d,p)//IEF-PCM-PBE0/6-311G(d,p) Approach[a]

subst	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp	ref
none	ethanol	250	377	321	280	271	294	322	4
2-Me	ethanol	252	380	324	282	272	296	324	4
1-Me	ethanol	259	399	339	291	280	306	331	4
2-Ome	ethanol	286	491	391	328	312	348	363	4
1,2-OMe	ethanol	279	482	386	323	307	341	374	4
1-Ome	ethanol	279	480	387	333	317	347	378	4
2-OH	ethanol	285	507	389	327	311	347	378	4
1-OH	ethanol	291	479	398	347	330	360	402	4
1,2,3-OH	ethanol	284	471	395	346	328	359	414	4
1,3-OH	ethanol	297	491	408	357	339	371	418	4,136
1,8-OH	ethanol	302	505	419	368	349	380	429	4
1,5-OH	ethanol	300	502	417	361	343	376	432	4
1,5-OH,3-Me	ethanol	299	498	415	360	342	374	433	137
1,2-OH	ethanol	299	524	426	359	340	376	435	4,136
1,3,8-OH,6-Me	ethanol	309	522	433	381	360	394	436	4
2-NH_2	ethanol	311	587	459	370	349	396	449	138
1,3,6,8-OH	ethanol	316	548	451	394	371	408	452	4
1-NH_2	ethanol	327	557	463	394	374	413	476	138
1,4-OH	ethanol	336	530	456	408	389	418	480	4
1,2,4-OH	ethanol	332	534	456	406	387	417	483	4
1,4,5-OH	ethanol	343	549	469	420	399	430	489	4
1,4,5,8-OH	ethanol	365	587	504	454	431	464	560	4
1,4-NH_2	ethanol	404	577	522	474	456	486	592	136
1,4-NHEt	ethanol	437	626	568	516	496	528	642	10

[a] PBE0 values are taken from ref 19. All values are in nm.

Table 2. λ_{max} (nm) of Azobenzenes (AB) Computed with the C-PCM-TD-X/6-311+G(2d,p)//C-PCM-PBE0/6-311G(d,p) Scheme

subst/compound	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp	ref
none	EtOH	297	377	342	310	300	325	320	139
4-F	EtOH	295	381	345	313	302	326	320	139
4-Me	EtOH	301	388	350	316	306	331	323	139
4-Br	EtOH	300	403	355	316	305	332	325	139
4,4′-F	EtOH	294	386	347	315	304	328	325	139
4,4′-Br	EtOH	303	425	367	322	310	339	326	139
4-phenylazomaleinanil	$CHCl_3$	302	406	357	318	308	334	329	2
4,4′-Me	EtOH	304	396	357	322	311	337	330	139
4-OMe	EtOH	306	414	366	329	318	343	344	136,139
4,4′-OMe	EtOH	312	433	380	341	329	356	355	136,139
4-NH_2	MeOH	330	445	399	360	348	375	386	2
4,4′-NH_2	EtOH	342	475	422	377	363	394	399	136
4-NMe_2	MeOH	334	477	418	371	357	387	408	2
4-NHPh	EtOH	335	524	438	372	357	394	411	2
6′-OBu,2,6-NH_2,3,3′-azopdipyridine	MeOH	330	442	401	377	366	383	435	2
2′-NH_2-azobenzenenaphthalene	MeOH	376	496	451	411	397	427	439	2
4-NO_2, 4′-NH_2	50% EtOH	365	603	483	402	389	428	443	2
2,4-NH_2-azobenzenenaphthalene	EtOH	338	476	441	381	370	386	451	2
4-NO_2, 4′-N(Et)(CH_2CH_2CN)	MeOH	358	629	492	399	384	428	455	2
4-NO_2, 4′-NHPh	50% EtOH	363	700	527	408	392	442	483	2
4-NO_2, 4′-N(Et)(CH_2CH_2OH)	50% EtOH	369	662	513	415	399	443	503	2
4-NO_2, 2-Cl,4′-N(Et)(CH_2CH_2OH)	50% EtOH	372	666	525	425	410	467	517	2

systematically improved by inclusion of LR terms. Finally, the molecular ordering is generally correct although one noteworthy mismatch could be detected (6′-OBu,2,6-NH_2, 3,3′-azopdipyridine versus 4,4′-NH_2): all approaches, but LC-ωPBE, disagree with experiments.

Table 3 summarizes our results for a 12 NQ set containing several 2,3-substituted structures that are known to be particularly problematic for global hybrids.[100] For these dyes, CAM-B3LYP outperforms PBE0, but the reverse is true for NQ with auxochroms at positions 5 and 8. One can also note

Table 3. First $\pi \rightarrow \pi^*$ Transition in 1,4-Naphthoquinone (NQ, C-PCM, Top) and Coumarins (CO, IEF-PCM, Bottom)[a]

subst	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp	ref
					1,4-Naphthoquinone				
2,3-Cl	methanol	258	412	349	296	283	316	337	140
2-OH	chcl$_3$	252	384	328	287	276	303	338	4,141
2,6-OH	ethanol	307	512	427	379	357	397	390	136
5-OMe	CHCl$_3$	281	521	411	350	330	368	396	4,141
2,3-Me,6-NH$_2$	cyclohexane	306	586	457	382	357	409	410	142
3,5-OH	CHCl$_3$	300	480	408	368	349	379	419	4,141
2,5-OH	CHCl$_3$	303	522	431	381	357	394	430	4,141
2,3-OH	CHCl$_3$	322	582	478	434	401	443	439	4,141
2-NHMe,3-Cl	cyclohexane	320	619	494	434	399	449	454	143
2-OMe,5,8-OH	CHCl$_3$	353	554	481	430	408	442	475	141
2-Cl,5,8-OH	CHCl$_3$	359	561	495	440	418	454	494	141
5-NH$_2$, 8-Ome	cyclohexane	361	599	515	464	438	479	564	142
					Coumarins				
4,7-OH (enol)	methanol	249	316	287	274	269	278	309	144
4-Me (enol)	ethanol	256	323	291	274	268	279	309	145,146
4-Me,7-OMe (enol)	water	260	340	303	285	278	290	322	147,148
4-Me,7-OH (enol)	water	261	337	303	284	277	288	323	147−150
7-OMe (enol)	water	266	341	306	290	283	295	324	151,152
4-Me,7-OH (cation)	water	274	374	330	307	301	312	345	147,149
4-Me,7-OMe (cation)	water	273	376	328	309	302	314	352	147
3-CN,7-OH (enol)	methanol	284	363	331	315	308	319	355	153
4-Me,7-OH (anion)	water	296	405	360	338	330	344	360	149,154,155
3-CN,7-OH (anion)	water	315	398	369	363	357	363	408	153

[a] All values are obtained through the PCM-TD-X/6-311+G(2d,p)//PCM-PBE0/6-311G(d,p) and are in nm. PBE0 values are from refs 99, 100, and 110 for NQ and CO, respectively.

Table 4. Diphenylamine (DPA, C-PCM, Top) and Diarylethenes (DA, IEF-PCM, Bottom) λ_{max} (nm), Obtained with the PCM-TD-X/6-311+G(2d,p)//PCM-PBE0/6-311G(d,p) Scheme[a]

		diphenylamine							
subst	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp	ref
none	hexane	248	320	284	273	265	280	286	156
4-NO$_2$	hexane	270	442	364	319	307	339	358	156−160
2,2′,4,4′-NO$_2$	ethanol	290−245	494−462	409−358	362−305	346−295	380−323	402−358	136,140,161
4,4′-NO$_2$	methanol	296	499	412	354	340	380	404	162
2,4′-NO$_2$	ethanol	296−258	532−455	434−364	370−310	352−300	397−331	407−353	161,163,164
2,4-NO$_2$	hexane	273−251	490−417	393−343	344−299	327−288	361−318	411−340	159
2-NO$_2$	hexane	289	497	416	364	346	384	415	156,158−160
2,2′-NO$_2$	hexane	286	509	417	362	343	383	420	158
2,6-NO$_2$	methanol	286	514	419	355	349	379	424	162

		diarylethenes							
R$_1$, R$_2$	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp.	ref
Cl, Cl	Hex	385	541	485	432	410	448	444	165,166
Cl, Ph	ACN	415	595	535	461	438	484	485	167
COOH,COOH	MeOH	440	700	599	509	477	537	531	165,166
Ph, Ph	Benz	456	687	605	507	479	537	531	165
COOEt,COOEt	ACN	429	670	578	496	465	520	540	167
p-Pyr., *p*-Pyr.	THF	468	744	636	525	494	559	551	168
CHO, CHO	Benz	471	791	670	563	522	592	580	165,166

[a] PBE0 figures are from refs 101 and 88 for DPA and DA, respectively.

systematic inversions (2,3-Cl versus 2-OH), but none of the proposed approach allows a λ_{max} classification clearly more appealing. Indeed, depending on the compounds, smaller (2,6-OH versus 5-OMe) or larger (2-NHMe,3-Cl versus 2-OMe,5,8-OH) proportion of exact exchange might help. Going from left to right in Table 3, we obtain MAE of 119, 99, 22, 42, 64, and 28 nm (the MSE are 119, −99, −11, 42, 64, and 26 nm). These values are in the line of the results obtained for AQ, although differences between PBE0 and CAM-B3LYP are strongly reduced for NQ.

The spectral data obtained for 10 neutral and charged CO are given in Table 3. The evaluation of the visible spectra

Table 5. Indigoids (IG) Main Visible Transition, Obtained with a IEF-PCM-TD-X/6-311+G(2d,p)//IEF-PCM-PBE0/6-311G(d,p) Scheme[a]

structure	subst	solvent	HF	PBE	PBE0	LC-PBE	LC-ωPBE	CAM-B3LYP	exp	ref
IG-a, X═X′═NH	none	CCl_4	400	651	572	519	490	525	594	169−172
	4,4′-Cl	$CHCl_3$	402	664	580	524	494	531	605	173,174
	4,4′-aza	EtOH	414	653	574	528	500	532	600	175
	5,5′-NO_2	TCE	383	637	550	503	475	507	580	176
	5,5′-Br	$CHCl_3$	407	708	599	534	502	542	611	173,174
	6,6′-NO_2	TCE	433	820	647	553	522	571	635	176
	6,6′-Br	TCE	399	636	569	519	490	524	588	176−178
	7,7′-Me	$CHCl_3$	406	671	586	527	498	534	612	174,179
	7,7′-aza	EtOH	370	624	539	490	461	494	556	175
X═X′═S	none	$CHCl_3$	354	655	544	465	428	483	546	180−187
	4,4′-Cl	benzene	356	646	546	467	431	484	548	188
	5,5′-NO_2	benzene	348	634	531	461	425	475	513	189
	5-OEt,5′-NO_2	benzene	367	841	603	487	447	512	562	190
	5-OEt,6′-NO_2	benzene	377	888	631	497	456	526	578	191
	6-NO_2	benzene	365	725	572	477	439	499	561	191
	7,7′-Br	benzene	353	663	546	464	428	481	546	192
X═X′═NMe	none	benzene	430	692	602	556	526	557	644	193
X═X′═O	none	CycloHex	292	524	430	379	358	386	413	194
X═X′═Se	none	benzene	354	647	568	471	433	496	562	195,196
X═NH, X′═S	none	benzene	376	657	556	488	456	501	574	197
IG-b	none	$CHCl_3$	339	560	494	448	415	454	505	185,186
IG-c	none	$CHCl_3$	327	529	456	418	387	424	458	185−187,198
IG-d	none	$CHCl_3$	323	550	472	427	392	434	467	187
IG-e	none	DMSO	324	438	395	366	353	374	413	134
IG-f	none	DMSO	346	582	490	438	415	446	516	134
IG-g	none	DMSO	329	556	444	371	357	391	455	134
IG-h, X═NH	none	DMSO	385	649	536	473	448	488	551	134
IG-h, X═S	none	Benzene	364	631	517	441	413	461	505	197
IG-i	none	DMSO	391	626	519	437	417	461	491	134
IG-j	none	methanol	256	368	325	298	290	305	355	199
IG-k	none	methanol	212	321	285	266	258	271	317	199

[a] More details can be found in refs 102−105.

of amino-CO is very challenging as both state-specific solvation and vibronic coupling could play a significant role.[112,113] Therefore, only hydroxy-CO are included in our set. As for the other dyes, HF (PBE) provides the smallest (largest) λ_{max} and hybrids stand in between, with wavelengths almost proportional to the exact exchange percentage at intermediate r_{12}. For CO, PBE, and PBE0 yield about the same accuracy but with opposite signed errors. Indeed the MSE are 67, −17, 20, 37, 43, and 33 nm for HF, PBE, PBE0, LC-PBE, LC-ωPBE, and CAM-B3LYP, respectively, while the MAE is equal to the MSE but for PBE (19 nm). All hybrids reproduce quite accurately the effect of protonation with predicted bathochromic shifts of 25 ± 2 nm to be compared to the experimental values of 22 (4-Me,7-OH−CO) and 30 nm (4-Me,7-OMe−CO). A basic medium modifies the λ_{max} of 3-CN,7-OH−CO by 53 nm, that is correctly reproduced by LC-PBE (48 nm) and LC-ωPBE (49 nm) but underestimated by HF (31 nm), PBE (35 nm), and PBE0 (38 nm). However, for 4-Me,7-OH CO, all hybrids exaggerate the impact of the enol-anion reaction.

Although DPA dyes show a significant charge-transfer character, global hybrids like PBE0 provide very accurate λ_{max} probably because the distance between the electron donor and the electron acceptor is rather small.[101] Table 4 confirms this conclusion with a MAE of 8 nm for PBE0 but 27 nm for CAM-B3LYP (other schemes produce even larger average errors). Some DPA present two peaks, and the only drawback of PBE0 is that the separation between these two absorptions ($\Delta\lambda$) is 70 nm instead of 54 nm for 2,4′-NO_2-DPA but 50 nm instead of 71 nm for 2,4-NO_2-DPA. However, this problem cannot be corrected by LR-DFT nor PBE nor HF. For the DA of Table 4, that are characterized by a very delocalized first excited state,[123] we have the reverse situation with a large error with PBE0 (MAE═64 nm), whereas LR-DFT are closer to the experimental data, especially CAM-B3LYP that provides a MAE of 8 nm and a maximal discrepancy limited to 20 nm.

Table 5 lists the results for more than 30 dyes of the IG family. For thioindigo (X═X′═S) derivatives, the PBE0 functional has been found astonishingly efficient[104] but is less accurate for indigo (X═X′═NH) dyes for which B3LYP was more adequate.[103] In the indirubin/isoindigo series (IG-e-IG-i), no global hybrid functional gives the correct ordering of the compounds.[105] In the IG-a series, modifying only the X and X′ atoms leads to the following λ_{max} ordering (in the nonpolar aprotic solvents used here): NMe,NMe > NH,NH > NH,S > Se,Se > S,S > O,O. This order is not reproduced by HF (that predicts no difference between thio and selenoindigo), nor PBE (that incorrectly foresees a bigger

λ_{max} for sulfur than amine-based compounds) nor PBE0 (that reverses the order of X=X′=Se and X=NH, X′=S), but is correctly predicted by the three range-separated functionals, a significant chemical success. Of course, for these six structures, the PBE0 wavelengths are always much closer to experiment than LR-DFT. However, relative changes are also better reproduced by LR-DFT. For instance, the wavelength difference between the selenoindigo and oxy-indigo is 3.0 times larger than the $\Delta\lambda_{max}$ separating indigo and its N−Me form (149/50 nm). This ratio is exaggerated by PBE0 (4.6) but reasonably reproduced by LC-PBE (2.5), LC-ωPBE (2.4), and CAM-B3LYP (2.8). Concerning the substitution of the outer-phenyl rings, all hybrid schemes provide the correct ordering for thioindigo and indigo but for some cases in which the experimental data are extremely close (within 1 or 2 nm). In the Wille and Lüttke "linkage" series,[134] all IG-e - IG-i, HF, PBE, PBE0, and CAM-B3LYP invert IG-i and IG-f, whereas LC-PBE and LC-ωPBE predict almost equal absorption wavelengths for these two dyes, that is certainly a major improvement. A couple of thioindigo structures[104] for which PBE0 produces the largest errors (5-OEt,5′-NO_2 and 5-OEt,6′-NO_2) are overcorrected by LR-DFT, and even CAM-B3LYP undershoots their λ_{max} by about 50 nm, confirming that the description of charge-transfer excited states is not systematically improved by range-separated functionals. Overall, the MAE are 170, 96, 19, 70, 99, and 58 nm for HF, PBE, PBE0, LC-PBE, LC-ωPBE, and CAM-B3LYP, respectively. This confirms the superiority of PBE0 for absolute wavelength estimations although LR-DFT appears to be authorized to give more consistent chemical conclusions.

B. Statistical Treatment and Corrections. Statistical analysis for the three main families (AQ, AB, and IG) and the complete set of dyes are given in Table 6. A graphical comparison is plotted in Figure 4 for the LC-ωPBE functional. It is clear that HF (PBE) considerably undershoots (overestimates) the experimental λ_{max} by 116 and 87 nm on average. Hybrids perform significantly better for all dye subsets (but CO). While CAM-B3LYP yields the smallest MSE and MAE for AB and DA, PBE0 appears more efficient for the other five families. On average, for the full set (118 transitions), we obtain a MAE of 22 nm only with PBE0, whereas CAM-B3LYP produces almost twice this error. Nevertheless, the PBE0 maximal deviations remain unacceptably large: −74 and +90 nm. In the eV scale, the PBE0 MAE attains 0.14 eV and the RMS is 0.17 eV. As we have considered a very extended and diverse set of dyes, this value can be regarded as a new 'expected accuracy' for organic dye design with TD-DFT. It is worth comparing this performance with the 0.19 eV (37 nm) MAE obtained with B3LYP/6-31G by Guillaumont and Nakamura for a smaller set of dyes. The two other extended studies are due to Fabian who reported B3LYP/6-31+G(d) MAE of 0.29 and 0.24 eV for $\pi \rightarrow \pi^*$ transitions in sulfur-free and sulfur-bearing molecules, respectively.[17,135] The increased performance here reported mainly originates in the use of (much) more extended basis sets and the explicit consideration of solvent effects, that are essential for a realistic simulation of the experimental setups.

Table 6. Statistical Analysis for the AQ, AB, and IG Series[a]

		without fit				with linear correction	
family	method	MSE	MAE	RMS	R^2	MAE	RMS
AQ	HF	127	127	132	0.97	11	14
	PBE	−71	74	80	0.78	28	36
	PBE0	12	19	27	0.96	12	15
	LC-PBE	67	67	71	0.98	8	10
	LC-ωPBE	85	85	89	0.99	8	9
	CAM-B3LYP	53	33	58	0.98	8	10
AB	HF	64	64	75	0.89	15	21
	PBE	−90	90	103	0.83	22	27
	PBE0	−20	25	27	0.93	12	17
	LC-PBE	33	33	43	0.95	10	14
	LC-ωPBE	46	46	54	0.96	10	14
	CAM-B3LYP	15	20	28	0.94	10	16
IG	HF	170	170	174	0.89	19	26
	PBE	−96	96	116	0.71	34	42
	PBE0	6	19	23	0.93	17	21
	LC-PBE	70	70	72	0.97	10	13
	LC-ωPBE	99	99	101	0.97	12	14
	CAM-B3LYP	58	58	60	0.98	10	12
All	HF	116	116	127	0.76	37	45
	PBE	−86	87	102	0.81	30	40
	PBE0	−3	22	29	0.91	22	28
	LC-PBE	52	52	58	0.94	18	23
	LC-ωPBE	71	71	78	0.94	18	23
	CAM-B3LYP	37	38	46	0.93	20	25

[a] All include the complete data from Tables 1−5. All values (but R^2) are given in nm.

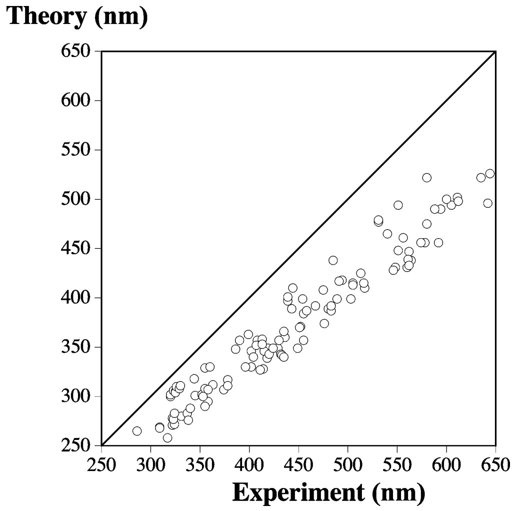


Figure 4. Comparison between the LC-ωPBE and measured λ_{max} (nm) for the full set of transitions. The central line indicates a perfect theory/experiment match.

To check the consistency between experimental and theoretical data, we have performed simple linear regressions on the different dye sets. Results are summarized in Table 6. HF and PBE obviously provide much smaller correlation coefficients than the hybrid approaches. Therefore, one can definitely discard HF and PBE for dye design: they provide not only the poorest absorption wavelength estimates but also the less consistent auxochromic displacements. Applying a linear correction to the PBE0 data is just useless, as the MAE

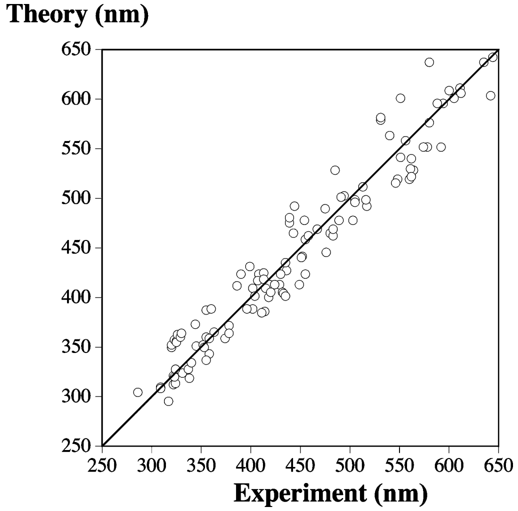


Figure 5. Comparison between experimental λ_{max} and the values obtained by eq 3.

and RMS are almost unmodified. On the contrary, the R^2 obtained with the range-separated functionals is at least 0.93, confirming the interest of such approaches for classifying molecules according to their transition energies. Therefore, a linear correction improves the results of the three TD-LR-DFT schemes, especially for the two LC functionals. For instance, using

$$\lambda^{\text{best}} = -38.63 + 1.295 \times \lambda^{\text{LC}-\omega\text{PBE}} \qquad (3)$$

provides a MAE limited to 18 nm and a rms of 23 nm, that are at least three times smaller than the uncorrected LC-ωPBE data. Additionally, this MAE represents a 20% improvement over the (raw or fitted) PBE0 error. The impact of eq 3 is illustrated in Figure 5, and it is striking that the maximal deviations are now limited to +57 and −41 nm, both being smaller than the prior-to-fitting MAE. From Table 6, it is also striking that considering a single dye family and performing a calibration is extremely efficient as a better correlation coefficient and smaller average errors are systematically attained with range-separated functionals. Therefore, if one is able to establish a calibration curve for a given family of dye, the use of TD-LR-DFT should lead to the sufficient accuracy for the design of a new dye structure.

IV. Conclusions

Using TD-DFT, we have assessed the efficiency of several functionals for reproducing the experimental UV/vis $\pi \rightarrow \pi^*$ absorption wavelength of a set of 100+ organic dyes belonging to the classes of major industrial interest: azobenzenes, anthraquinones, indigos, diarylethenes, ... It was found that the computed λ_{max} systematically obey a PBE > PBE0 > CAM-B3LYP > LC-PBE > LC-ωPBE > HF order. This result can be rationalized by the total amount of exact exchange in each functional. Overall, PBE0 provides the smallest error with an average absolute deviation limited to 0.14 eV/22 nm. We state that this value should be regarded as a reference *expected PCM-TD-PBE0 accuracy* for low-lying excited states of conjugated organic compounds. The second best approach, CAM-B3LYP, suffers larger deviations (0.26 eV/38 nm) but appears particularly well suited for studying dyes with a very delocalized excited state. On the contrary, HF and PBE give very poor estimates with average errors of 0.97 eV/116 nm and 0.45 eV/87 nm, respectively. If range-separated hybrids cannot beat PBE0 in terms of absolute λ_{max}, they provide more consistent evaluations of the auxochromic shifts. Indeed, linear fittings demonstrate that LR-functionals systematically give large R^2. Consequently, using a calibration equation such as eq 3 can considerably improve the accuracy of the LR computations. This is especially true when the statistical treatment is performed for a given dye family, for which average errors close to 10 nm are indeed obtained, allowing an efficient dye molecular design as it combines nearly quantitative wavelength estimates to chemically sound classifications. For the complete set of dyes, using scaled LR-DFT improves the PBE0 MAE by about 20%.

Of course, only low-lying π^* excited states of conjugated organic molecules have been considered in the present investigation. Care should be taken before applying eq 3 to other types of excitations or structures. However, for Rydberg states in small molecules it is clear that range-separated hybrids are adequate,[64,73,85] whereas these functionals are as accurate as global hybrids for $n \rightarrow \pi^*$ transitions.[88] Therefore, this contribution paves the way toward accurate, yet affordable, estimations of the excited-state energies of medium and large molecules. We are currently investigating inorganic structures to test the transferability of this approach.

Acknowledgment. D.J. and E.A.P. thank the Belgian National Fund for their research associate positions. D.J., E.A.P., I.C., and C.A. thank the Commissariat Général aux Relations Internationales (CGRI) and the Egide agency for supporting this work within the framework of the Tournesol Scientific cooperation between the Communauté Française de Belgique, and France. HF, PBE, and PBE0 calculations have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the "Loterie Nationale" for the convention number 2.4578.02 and of the FUNDP. The work at Rice University was supported by NSF (CHE-0457030) and the Welch Foundation.

References

(1) Griffiths, J. *Colour and Constitution of Organic Molecules;* Academic Press: London, 1976.

(2) Green, F. J. *The Sigma-Aldrich Handbook of Stains, Dyes and Indicators;* Aldrich Chemical Company, Inc.: Milwaukee, WI, 1990.

(3) Zollinger, H. *Color Chemistry, Syntheses, Properties and Applications of Organic Dyes and Pigments,* 3rd ed.; Wiley-VCH: Weinheim, Germany, 2003.

(4) Thomson, R. H. *Naturally Occurring Quinones,* 2nd ed.; Academic Press: London, U.K., 1971.

(5) Natansohn, A.; Rochon, P. *Chem. Rev.* **2002**, *102*, 4139−4176.

(6) Murakami, H.; Kawabuchi, A.; Kotoo, K.; Kunitake, M.; Nakashima, N. *J. Am. Chem. Soc.* **1997**, *119*, 7605−7606.

(7) Qu, D. H.; Wang, Q. C.; Ren, J.; Tian, H. *Org. Lett.* **2004**, *6*, 2085−2088.

(8) Fry, M.; Pudney, M. *Biochem. Pharmacol.* **1992**, *43*, 1545−1553.

(9) Cai, P.; Kong, F.; Ruppen, M.; Glasier, G.; Carter, G. *J. Nat. Prod.* **2005**, *68*, 1736−1742.

(10) Christie, R. M. *Colour Chemistry;* The Royal Society of Chemistry: Cambridge, U.K., 1991.

(11) Michaux, C.; Rolin, S.; Dogne, J.-M.; Durant, F.; Masereel, B.; Delarge, J.; Wouters, J. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1019−1022.

(12) Michaux, C.; Charlier, C.; Julemont, F.; de Leval, X.; Dogne, J. M.; Pirotte, B.; Durant, F. *Eur. J. Med. Chem.* **2005**, *40*, 1316−1324.

(13) Irie, M. *Chem. Rev.* **2000**, *100*, 1685−1716.

(14) Feringa, B. L. *Molecular Switches;* Wiley-VCH: Weinheim, Germany, 2001.

(15) Perrier, A.; Maurel, F.; Aubard, J. *J. Photochem. Photobiol. A: Chem.* **2007**, *189*, 167−176.

(16) Schäfer, A. *Modern Methods and Algorithms of Quantum Chemistry;* 2nd ed.; John von Neumann Institute for Computing: Jülich, Germany, 2000; Vol. 3 of *NIC*.

(17) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199−217.

(18) Jacquemin, D.; Perpete, E. A. *Chem. Phys. Lett.* **2006**, *429*, 147−152.

(19) Jacquemin, D.; Assfeld, X.; Preat, J.; Perpete, E. A. *Mol. Phys.* **2007**, *105*, 325−331.

(20) Perdew, J. P.; Ruzsinsky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 06 2001.

(21) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708−4717.

(22) Jacquemin, D.; Preat, J.; Wathelet, V.; André, J. M.; Perpete, E. A. *Chem. Phys. Lett.* **2005**, *405*, 429−433.

(23) Jacquemin, D.; Preat, J.; Perpete, E. A. *Chem. Phys. Lett.* **2005**, *410*, 254−259.

(24) Petit, L.; Quartarolo, A.; Adamo, C.; Russo, N. *J. Phys. Chem. B* **2006**, *110*, 2398−2404.

(25) Quartarolo, A. D.; Russo, N.; Sicilia, E. *Chem. Eur. J.* **2006**, *12*, 6797−6803.

(26) Petit, L.; Adamo, C.; Russo, N. *J. Phys. Chem. B* **2005**, *109*, 12214−12221.

(27) Quartarolo, A. D.; Russo, N.; Sicilia, E.; Lelj, F. *J. Chem. Theory Comput.* **2007**, *3*, 860−869.

(28) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(29) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(30) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664−675.

(31) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264−6271.

(32) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624−9631.

(33) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029−5036.

(34) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(35) Hoe, W. M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319−328.

(36) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559−9569.

(37) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129−12137.

(38) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405−3416.

(39) Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673−2677.

(40) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

(41) Paizs, B.; Suhai, S. *J. Comput. Chem.* **1998**, *19*, 575−584.

(42) Jacquemin, D.; André, J. M.; Perpete, E. A. *J. Chem. Phys.* **2004**, *121*, 4389−4396.

(43) Jacquemin, D.; Femenias, A.; Chermette, H.; Ciofini, I.; Adamo, C.; André, J. M.; Perpete, E. A. *J. Phys. Chem. A* **2006**, *110*, 5952−5959.

(44) Champagne, B.; Perpete, E. A.; van Gisbergen, S.; Baerends, E. J.; Snijders, J. G.; Soubra-Ghaoui, C.; Robins, K.; Kirtman, B. *J. Chem. Phys.* **1998**, *109*, 10489−10498.

(45) Champagne, B.; Perpete, E. A.; Jacquemin, D.; van Gisbergen, S.; Baerends, E.; Soubra-Ghaoui, C.; Robins, K.; Kirtman, B. *J. Phys. Chem. A* **2000**, *104*, 4755−4763.

(46) Guillaumont, D.; Nakamura, S. *Dyes Pigm.* **2000**, *46*, 85−92.

(47) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697−12699.

(48) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007−4016.

(49) Magyar, R. J.; Tretiak, S. *J. Chem. Theory Comput.* **2007**, *3*, 976−987.

(50) Krieger, J. B.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *45*, 101−126.

(51) Legrand, C.; Suraud, E.; Reinhard, P. G. *J. Phys. B* **2002**, *35*, 1115−1128.

(52) Ciofini, I.; Chermette, H.; Adamo, C. *Chem. Phys. Lett.* **2003**, *380*, 12−20.

(53) Vignale, G.; Kohn, W. *Phys. Rev. Lett.* **1996**, *77*, 2037−2040.

(54) van Faasen, M.; Boeij, P. L.; van Leeuwen, R.; Berger, J. A.; Snijders, J. G. *Phys. Rev. Lett.* **2002**, *88*, 186401.

(55) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287−5293.

(56) Alonso, J. A.; Mananes, A. *Theor. Chem. Acc.* **2007**, *117*, 467−472.

(57) Wang, D. Y.; Ye, Q.; Li, B. G.; Zhang, G. L. *Nat. Prod. Res.* **2003**, *17*, 365−368.

(58) Bulat, F. A.; Toro-Labbé, A.; Champagne, B.; Kirtman, B.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 014319.

(59) Savin, A. In *Recent Developments and Applications of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996; Chapter 9, pp 327−354.

(60) Leininger, T.; Stoll, H.; Werner, H. J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151−160.

(61) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540−3544.

(62) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207−8215; **2006**, *124*, 219906 (E).

(63) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.

(64) Tawada, T.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425−8433.

(65) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51−56.

(66) Yanai, T.; Harrison, R. J.; Handy, N. C. *Mol. Phys.* **2005**, *103*, 413−424.

(67) Kamiya, M.; Sekino, H.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *122*, 234111.

(68) Rudberg, E.; Salek, P.; Helgaker, T.; Agren, H. *J. Chem. Phys.* **2005**, *123*, 184108.

(69) Sato, T.; Tsuneda, T.; Hirao, K. *Mol. Phys.* **2005**, *103*, 1151−1164.

(70) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *123*, 104307.

(71) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.

(72) Peach, M. J. G.; Helgaker, T.; Salek, P.; Keal, T. W.; Lutnaes, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558−562.

(73) Peach, M. J. G.; Cohen, A. J.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4543−4549.

(74) Chiba, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2006**, *124*, 144106.

(75) Cai, Z. L.; Crossley, M. J.; Reimers, J. R.; Kobayashi, R.; Amos, R. D. *J. Phys. Chem. B* **2006**, *110*, 15624−15632.

(76) Kobayashi, R.; Amos, R. D. *Chem. Phys. Lett.* **2006**, *420*, 106−109.

(77) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(78) Vydrov, O. A.; Heyd, J.; Krukau, V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(79) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 154109.

(80) Sekino, H.; Maeda, Y.; Kamiya, M.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 014107.

(81) Chiba, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 034504.

(82) Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(83) Jacquemin, D.; Perpete, E. A.; Medved', M.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 191108.

(84) Nguyen, K. A.; Day, P. N.; Pachter, R. *J. Chem. Phys.* **2007**, *126*, 094303.

(85) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932−2941.

(86) Fromager, E.; Toulouse, J.; Jensen, H. J. A. *J. Chem. Phys.* **2007**, *126*, 074111.

(87) Sekino, H.; Maeda, Y.; Kamiya, M. *Mol. Phys.* **2005**, *103*, 2183−2189.

(88) Jacquemin, D.; Perpete, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 094102.

(89) Jacquemin, D.; Preat, J.; Charlot, M.; Wathelet, V.; André, J. M.; Perpete, E. A. *J. Chem. Phys.* **2004**, *121*, 1736−1743.

(90) Perpete, E. A.; Wathelet, V.; Preat, J.; Lambert, C.; Jacquemin, D. *J. Chem. Theory Comput.* **2006**, *2*, 434−440.

(91) Shen, L.; Ji, H. F.; Zhang, H. Y. *J. Mol. Struct. (THEOCHEM)* **2006**, *758*, 221−224.

(92) Zhang, Q. M.; Gong, X. D.; Xiao, H. M.; Xu, X. J. *Acta Chim. Sin.* **2006**, *64*, 381−387.

(93) Jacquemin, D.; Wathelet, V.; Preat, J.; Perpete, E. A. *Spectrochim. Acta, Part A* **2007**, 67, 334−341.

(94) Fliegl, H.; Köhn, A.; Hättig, C.; Ahlrichs, R. *J. Am. Chem. Soc.* **2003**, *125*, 9821−9827.

(95) Chen, P. C.; Chieh, Y. C.; Wu, J. C. *J. Mol. Struct. (THEOCHEM)* **2005**, *715*, 183−189.

(96) Liu, J. N. *J. Mol. Struct. (THEOCHEM)* **2005**, *730*, 151−154.

(97) Briquet, L.; Vercauteren, D. P.; Perpete, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2006**, *417*, 190−195.

(98) Briquet, L.; Vercauteren, D. P.; André, J. M.; Perpete, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2007**, *435*, 257−262.

(99) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpete, E. A. *Chem. Phys.* **2006**, *328*, 324−332.

(100) Perpete, E. A.; Lambert, C.; Wathelet, V.; Preat, J.; Jacquemin, D. *Spectrochim. Acta, Part A* **2007**, *68*, 1326−1333.

(101) Jacquemin, D.; Bouhy, M.; Perpete, E. A. *J. Chem. Phys.* **2006**, *124*, 204321.

(102) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpete, E. A. *J. Mol. Struct. (THEOCHEM)* **2005**, *731*, 67−72.

(103) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpete, E. A. *J. Chem. Phys.* **2006**, *124*, 074104.

(104) Jacquemin, D.; Preat, J.; Wathelet, V.; Fontaine, M.; Perpete, E. A. *J. Am. Chem. Soc.* **2006**, *128*, 2072−2083.

(105) Perpete, E. A.; Preat, J.; André, J. M.; Jacquemin, D. *J. Phys. Chem. A* **2006**, *110*, 5629−5635.

(106) Cave, R. J.; Castner, E. W., Jr. *J. Phys. Chem. A* **2002**, *106*, 12117−12123.

(107) Cave, R. J.; Burke, K.; Castner, E. W., Jr. *J. Phys. Chem. A* **2002**, *106*, 9294−9305.

(108) Preat, J.; Jacquemin, D.; Perpete, E. A. *Chem. Phys. Lett.* **2005**, *415*, 20−24.

(109) Preat, J.; Jacquemin, D.; Wathelet, V.; André, J. M.; Perpete, E. A. *J. Phys. Chem. A* **2006**, *110*, 8144−8150.

(110) Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Assfeld, X.; Ciofini, I.; Adamo, C. *J. Chem. Phys.* **2006**, *125*, 164324.

(111) Santoro, F.; Improta, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 084509.

(112) Improta, R.; Barone, V.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 054103.

(113) Improta, R.; Barone, V.; Santoro, F. *Angew. Chem., Int. Ed.* **2007**, *46*, 405−408.

(114) Majumdar, D.; Lee, H. M.; Kim, J.; Kim, K. S.; Mhin, B. J. *J. Chem. Phys.* **1999**, *111*, 5866−5872.

(115) Kobatake, S.; Morimoto, M.; Asano, Y.; Murakami, A.; Nakamura, S.; Irie, M. *Chem. Lett.* **2002**, 1224−1225.

(116) Goldberg, A.; Murakami, A.; Kanda, K.; Kobayashi, T.; Nakamura, S.; Ucjida, K.; Sekiya, H.; Fukaminato, T.; Kawau, T.; Kobatake, S.; Irie, M. *J. Phys. Chem. A* **2003**, *107*, 4982−4988.

(117) Giraud, M.; Léaustic, A.; Charlot, M. F.; Yu, P.; Césario, M.; Philouze, C.; Pansu, R.; Nakatani, K.; Ishow, E. *New. J. Chem.* **2005**, *29*, 439−446.

(118) Higashiguchi, K.; Matsuda, K.; Asano, A.; Murakami, A.; Nakamura, S.; Irie, M. *Eur. J. Org. Chem.* **2005**, 91−97.

(119) Clark, A. E. *J. Phys. Chem. A* **2006**, *110*, 3790−3796.

(120) Chen, D. Z.; Wang, Z.; Zhao, X. *J. Mol. Struct. (THEOCHEM)* **2006**, *774*, 77−81.

(121) Yokojima, S.; Matsuda, K.; Irie, M.; Murakami, A.; Kobayashi, T.; Nakamura, S. *J. Phys. Chem. A* **2006**, *110*, 8137−8143.

(122) Perpete, E. A.; Jacquemin, D. *J. Photochem. Photobiol. A: Chem.* **2007**, *187*, 40−44.

(123) Perpete, E. A.; Maurel, D.; Jacquemin, D. *J. Phys. Chem. A* **2007**, *111*, 5528−5535.

(124) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(125) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274−7280.

(126) Song, J. W.; Hirosawa, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 154105.

(127) 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.

(128) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Kudin, K. N.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; 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.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Brothers, E.; Staroverov, V.; Kobayashi, R.; Normand, J.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian DVP, Revision C.01*; Gaussian, Inc.: Wallingford, CT, 2006.

(129) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218−8224.

(130) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889−2899.

(131) Curtiss, L. A.; Raghavachari, K.; Referm, P. C.; Pople, J. A. *Chem. Phys. Lett.* **1997**, *270*, 419−426.

(132) Barone, V.; Adamo, C. *J. Chem. Phys.* **1996**, *105*, 11007−11019.

(133) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999−3094.

(134) Wille, E.; Lüttke, W. *Chem. Ber.* **1973**, *106*, 3240−3257.

(135) Fabian, J.; Diaz, L. A.; Seifert, G.; Niehaus, T. *J. Mol. Struct. (THEOCHEM)* **2002**, 594, 41−53.

(136) Weast, R. C. *Handbook of Chemistry and Physics*, 51st ed.; The Chemical Rubber Company: Cleveland, OH, 1970.

(137) Günaydin, K.; Topcu, G.; Ion, R. M. *Nat. Prod. Lett.* **2002**, *16*, 65−70.

(138) Fabian, J.; Nepras, M. *Collect. Czech. Chem. Commun.* **1980**, 45, 2605−2620.

(139) Gore, P. H.; Wheeler, O. H. *J. Org. Chem.* **1961**, *26*, 3295−3298.

(140) Grasselli, J. G. *Atlas of Spectral Data and Physical Constants for Organic Compounds;* The Chemical Rubber Company: Cleveland, OH, 1973.

(141) Singh, I.; Ogata, R. T.; Moore, R. E.; Chang, C. W. J.; Scheuer, P. J. *Tetrahedron* **1968**, *24*, 6053−6073.

(142) Chu, K. Y.; Griffiths, J. *J. Chem. Res.* **1978**, 180−181.

(143) Chu, K. Y.; Griffiths, J. *J. Chem. Soc., Perkin Trans. I* **1978**, 1083−1087.

(144) Wolfbeis, O. S.; Uray, G. *Monat. Chem.* **1978**, *109*, 123−136.

(145) Ganguly, B. K.; Bagchi, P. *J. Org. Chem.* **1956**, *21*, 1415−1419.

(146) Wheelcock, C. E. *J. Am. Chem. Soc.* **1959**, *81*, 1348−1352.

(147) Yakatan, G. J.; Juneau, R. J.; Schulman, S. G. *Anal. Chem.* **1972**, *44*, 1044−1046.

(148) Giri, R.; Rathi, S. S.; Machwe, M. K.; Murti, V. V. S. *Spectrochim. Acta, Part A* **1988**, *44*, 805−807.

(149) Moriya, T. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 1873−1886.

(150) Kumar, S.; Rao, V. C.; Rastogi, R. C. *Spectrochim. Acta, Part A* **2001**, *57*, 41−47.

(151) Heldt, J. R.; Helds, J.; Ston, M.; Diehl, H. A. *Spectrochim. Acta, Part A* **1995**, *51*, 1549−1563.

(152) Azuma, K.; Suzuki, S.; Uchiyama, S.; Kajiro, T.; Santa, T.; Imai, K. *Photochem. Photobiol. Sci.* **2003**, *2*, 443−449.

(153) Wolfbeis, O. S. *Z. Naturforsch., Teil A* **1977**, *32A*, 1065−1067.

(154) Sherman, W. R. *Anal. Chem.* **1968**, *40*, 803−805.

(155) Sun, W. C.; Gee, K. R.; Haugland, R. P. *Bioorg. Med Chem. Lett.* **1998**, *8*, 3107−3110.

(156) Lutskii, A. E.; Konel'skaya, V. N. *Z. Obs. Khim.* **1960**, *30*, 3773−3782.

(157) Bugai, P. M.; Konel'skaya, V. N. *Izv. Akad. Nauk. SSSR - Ser. Fiz.* **1954**, *18*, 695−697.

(158) Bugai, P. M.; Konel'skaya, V. N.; Gol'berkova, A. S.; Bazhenova, L. M. *Z. Fiz. Khim.* **1962**, *36*, 2233−2234.

(159) Bell, M. G. W.; Day, M.; Peters, A. T. *J. Soc. Dyers Colour.* **1966**, *82*, 410−414.

(160) Day, M.; Peters, A. T. *J. Soc. Dyers Colour.* **1969**, *85*, 8−13.

(161) Schroeder, W. A.; Wilcox, P. E.; Trueblood, K. N.; Dekker, A. O. *Anal. Chem.* **1951**, *23*, 1740−1747.

(162) Csaszar, J. *Acta Phys. Chem.: (Szeged)* **1987**, *33*, 11−21.

(163) Asquith, R. S.; Bridgeman, I.; Peters, A. T. *J. Soc. Dyers Colour.* **1965**, *81*, 439−441.

(164) Asquith, R. S.; Peters, A. T.; Wallace, F. *J. Soc. Dyers Colour.* **1968**, *84*, 507−510.

(165) Lucas, L. N. Dithienylcyclopentene Optical Switches: Towards Photoresponsive Supramolecular Materials, Ph.D. thesis, Rijksuniversiteit Groningen: 2001.

(166) Lucas, L. N.; de Jong, J. J.; van Esch, J. H.; Kellogg, R. M.; Feringa, B. L. *Eur. J. Org. Chem.* **2003**, 155−166.

(167) Browne, W. R.; de Jong, J. J.; Kudernac, T.; Walko, M.; Lucas, L. N.; Uchida, K.; van Esch, J. H.; Feringa, B. L. *Chem. Eur. J.* **2005**, *11*, 6430−6441.

(168) Qin, B.; Yao, R.; Zhao, X.; Tian, H. *Org. Biolmol. Chem.* **2003**, *1*, 2187−2191.

(169) Sheppard, S. E.; Newsome, P. T. *J. Am. Chem. Soc.* **1942**, *64*, 2937−3946.

(170) Klessinger, M.; Lüttke, W. *Tetrahedron* **1963**, 19 Suppl. 2, 315−335.

(171) Klessinger, M.; Lüttke, W. *Chem. Ber.* **1966**, *99*, 2136−2145.

(172) Monahan, A. R.; Kuder, J. E. *J. Org. Chem.* **1972**, *37*, 4182−4184.

(173) Brode, W. R.; Pearson, E. G.; Wyman, G. M. *J. Am. Chem. Soc.* **1954**, *76*, 1034−1036.

(174) Weinstein, J.; Wyman, G. M. *J. Am. Chem. Soc.* **1956**, *78*, 2387−2390.

(175) Lüttke, W.; Hunsdiecker, D. *Chem. Ber.* **1966**, *99*, 2146−2154.

(176) Saddler, P. W. *J. Org. Chem.* **1956**, *21*, 316−318.

(177) Friedländer, P.; Bruckner, S.; Deutsch, G. *J. Liebigs Ann. Chem.* **1912**, *388*, 23−49.

(178) Ettinger, L.; Friedländer, P. A. *Ber. Dtsch. Chem. Ges.* **1912**, *45*, 2074−2080.

(179) Travasso, M. I. G.; Santos, P. C. S.; Oliveira-Campos, A. M. F.; Raposo, M. M. M.; Prasitpan, N. *Adv. Colour Sci. Technol.* **2003**, *6*, 95−99.

(180) Friedländer, P. *Ber. Dtsch. Chem. Ges.* **1908**, *41*, 772−777.

(181) Wyman, G. M.; Brode, W. R. *J. Am. Chem. Soc.* **1951**, *73*, 1487−1493.

(182) Brode, W. R.; Wyman, G. M. *J. Am. Chem. Soc.* **1951**, *73*, 4267−4270.

(183) Pummerer, R.; Marondel, G. *Chem. Ber.* **1960**, *99*, 2834−2839.

(184) Mostoslavskii, M. A.; Izmail'ski, V.; Shapkina, M. M. *J. Gen. Chem. USSR* **1962**, 32, 1731−1739.

(185) Lüttke, W.; Hermann, H.; Klessinger, M. *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 598−599.

(186) Hermann, H.; Lüttke, W. *Chem. Ber.* **1968**, *101*, 1715−1728.

(187) Luhmann, U.; Wentz, F. G.; Knieriem, B.; Lüttke, W. *Chem. Ber.* **1978**, *111*, 3233−3245.

(188) Kirsch, A. D.; Wyman, G. M. *J. Phys. Chem.* **1977**, *81*, 413−420.

(189) Dokunikhin, N. S.; Gerasimenko, Y. E. *J. Gen. Chem. USSR* **1960**, *30*, 655−658.

(190) Dokunikhin, N. S.; Gerasimenko, Y. E. *J. Gen. Chem. USSR* **1960**, *30*, 1966−1968.

(191) Dokunikhin, N. S.; Gerasimenko, Y. E. *J. Gen. Chem. USSR* **1961**, *31*, 205−208.

(192) Dokunikhin, N. S.; Gerasimenko, Y. E. *J. Gen. Chem. USSR* **1960**, *30*, 1253−1255.

(193) Giuliano, C. R.; Hess, L. D.; Margerum, J. D. *J. Am. Chem. Soc.* **1968**, *90*, 587−594.

(194) Gusten, H. *Chem. Commun.* **1969**, 133−134.

(195) Fitjer, L.; Lüttke, W. *Chem. Ber.* **1972**, *105*, 919−928.

(196) Wyman, G. M.; Zarnegar, B. *J. Phys. Chem.* **1973**, *73*, 831−837.

(197) Haucke, G.; Paetzold, R. *Nova Acta Leopold. Suppl.* **1978**, *11*, 1−123.

(198) Gerken, R.; Fitjer, L.; Müller, P.; Usón, I.; Kowski, K.; Rademacher, P. *Tetrahedron* **1999**, *55*, 14429−14434.

(199) Junek, H.; Fischer-Colbrie, H.; Sterk, H. *Chem. Ber.* **1977**, *110*, 2276−2285.

CT700187Z

JCTC Journal of Chemical Theory and Computation

Application of the TraPPE Force Field for Predicting the Hildebrand Solubility Parameters of Organic Solvents and Monomer Units

Neeraj Rai,[†] Alexander J. Wagner,[‡] Richard B. Ross,[‡] and J. Ilja Siepmann*,[†]

Departments of Chemistry and of Chemical Engineering and Material Science, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota 55455, and Corporate Research Materials Laboratory, 201-2E-23, 3M Company, St. Paul, Minnesota 55144

Received June 4, 2007

Abstract: Configurational-bias Monte Carlo simulations in the isothermal−isobaric and Gibbs ensembles using the transferable potentials for phase equilibria (TraPPE) force field were carried out to compute the liquid densities, the Hildebrand solubility parameters, and the heats of vaporization for a set of 32 organic molecules with different functional groups at a temperature of 298.15 K. In addition, the heats of vaporization were determined at the normal boiling points of these compounds. Comparison to experimental data demonstrates that the TraPPE force field is significantly more accurate than predictions obtained from molecular dynamics simulations with the Dreiding force field [Belmares et al. *J. Comput. Chem.* **2004**, *25*, 1814] and an equation of state approach [Stefanis et al. *Fluid Phase Equil.* **2006**, *240,* 144]. For the TraPPE force field, the mean unsigned percent errors for liquid density, the Hildebrand solubility parameter, and the heat of vaporization at 298.15 K are 1.3, 3.3, and 4.5%, respectively.

1. Introduction

The solubility parameter, δ_H, proposed by Hildebrand based on regular solution theory,[1] is used frequently to predict miscibility behavior for technological applications, such as the blending of oil fractions to meet end product specifications,[2] the prevention of asphaltene precipitation,[3] the estimation of the shelf life of polymers and drug formulations,[4−7] the development of synthetic membranes,[8] self-assembly and gelation processes,[9] and the formation of micelles[10] and nanocomposites.[11] In addition, numerous group contribution based techniques to predict and correlate polymer properties such as the glass transition temperature and the permeability of molecules through membranes depend on accurate estimates of the solubility parameter.[12] As the solubility parameter is obtained from the cohesive energy density, ρ_U, it is a measure of the interactions among the molecules in the condensed phase.[1] The overall interactions between molecules, as a first approximation, can be thought of as the sum of dispersion and first-order electrostatic interactions. Along similar lines, Hansen proposed a three-component (dispersion, polar, and hydrogen bonding) solubility parameter model.[13] However, while the cohesive energy density can be measured for certain systems, it is not possible to cleanly separate the individual contributions from dispersion, polar, and hydrogen-bonding interactions.[12,14]

For low-molecular-weight solvents and monomer units, direct experimental measurements of the heat of vaporization, ΔH_{vap}, and molar volume can be used to determine the cohesive energy density and, hence, the Hildebrand solubility parameter. On the other hand, for natural and synthetic polymers and other high-molecular-weight compounds, for which the determination of the heat of vaporization is impractical due to extremely low vapor pressures at room temperature and chemical degradation at elevated temperatures, miscibility experiments are often carried out to deduce ρ_U and δ_H through comparison with compounds with known δ_H.[15] Other indirect methods to estimate δ_H include gas chromatography[16−19] and transport and mechanical properties,

* Corresponding author e−mail: siepmann@chem.umn.edu.
[†] University of Minnesota.
[‡] 3M Company.

10.1021/ct700135j CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/29/2007

such as the viscosity[7] and Young's modulus.[20] More recently, computational tools ranging from quantitative structure−property relationships (QSPR)[12,21,22] over equation of state approaches[23] to atomistic molecular dynamics (MD) and Monte Carlo (MC) simulations[24−30] have been applied for the correlation and prediction of δ_H. Although molecular simulations are computationally much more expensive than QSPR calculations, the former rely less on experimental data and can provide molecular-level information in addition to thermodynamic quantities.

Already in 1985, Theodorou and Suter[24] used an atomistic model to calculate the cohesive energy density and the Hildebrand solubility parameter of atactic polypropylene. Only 15 configurations were used to compute the averages, and δ_H was found to be underestimated by approximately 15%. A few years later, Choi et al.[25] used MD simulations to calculate the three-dimensional solubility parameters of alkyl phenol ethoxylates and reported good agreement with estimates from group contribution models.[31] These authors also explored the effect of different schemes to select partial atomic charges and found that the polar component of the Hansen solubility parameter can differ significantly (3 hildebrands). Lago et al.[26] used MC simulations in the Gibbs ensemble to compute the solubility parameter for organic solvents and diatomics. Recently, MD simulations were carried out to investigate the binary blend compatibility of poly(vinyl alcohol) and poly(methyl methacrylate)[30] and various surface properties and solubility parameters for perfluorinated homopolymers and their random copolymers.[32] Closely related to the aim of the present work, Belmares et al.[29] used the Dreiding force field[31] along with charges derived from molecular electrostatic potentials (ESP) or Mulliken population analysis to calculate solubility parameters for a large set of common organic solvents and monomer units. These authors found regression and correlation coefficients of 1.01 and 0.73, respectively, when comparing the predicted δ_H with experimental data, indicating a relatively large scatter in predicted values.

As the solubility parameter is related to vapor−liquid equilibrium properties, the present work attempts to assess whether modern force fields parametrized to phase equilibrium data can be used successfully for the prediction of δ_H not only at room temperature but also at elevated temperatures. To this extent we employ the transferable potentials for phase equilibria (TraPPE) force field that has been used extensively to predict fluid phase equilibria,[33−35] retention in chromatography,[36−39] octanol−water partitioning,[40] adsorption in pharmaceutical solids,[41] and transport properties.[42−45] A brief outline of this article is as follows: section 2 provides a short definition of the solubility parameter and descriptions of the simulation methodology and force field. In section 3, the predictions of δ_H, ρ_U, ΔH_{vap}, and the specific densities made using the TraPPE force field are compared with values obtained from experiments, MD simulations with the Dreiding force field,[29,31] and an equation of state approach.[23]

2. Methodology and Simulation Details

2.1. Background. The cohesive energy of a condensed phase corresponds to the increase in internal energy when all the intermolecular interactions are eliminated per mole of condensed phase of a substance.[12,21] At a particular temperature, T, and the corresponding saturation pressure, p_{sat}, the cohesive energy density, ρ_U, is obtained by dividing the cohesive energy by the molar volume of the condensed phase

$$\rho_U(T,P_{sat}) = \frac{U_{coh}(T,P_{sat})}{V_{liq}(T,P_{sat})} \quad (1)$$

where U_{coh} and V_{liq} are the molar cohesive energy and the molar volume of the liquid phase, respectively. For low-molecular-weight compounds, the cohesive energy is obtained from the molar heat of vaporization, using the following equation

$$U_{coh}(T, P_{sat}) = \Delta H_{vap}(T) - P_{sat}\Delta V \quad (2)$$

where ΔH_{vap} and ΔV are the enthalpy of vaporization and the difference in vapor and liquid molar volumes, respectively. If the vapor pressure of a substance is negligible and if there is no aggregation of molecules in the vapor phase, then the vapor phase behaves like ideal gas. In such cases, eq 2 can be simplified as

$$U_{coh}(T, P_{sat}) = \Delta H_{vap} - RT \quad (3)$$

where R and T are the universal gas constant and the absolute temperature, respectively. The Hildebrand solubility parameter, δ_H, is calculated directly from ρ_U using the equation

$$\delta_H = (\rho_U)^{1/2} \quad (4)$$

As ρ_U and the pressure have the same dimension, its SI unit and that for δ_H are Pascal (Pa) and $\mathrm{Pa}^{1/2}$, respectively. Traditionally, the solubility parameter has been expressed in $(\mathrm{cal/cm}^3)^{1/2}$, called a "hildebrand''. In this work, the "hildebrand'' is used as the unit for the solubility parameter.

2.2. Simulation Details. Coupled−decoupled configurational-bias Monte Carlo (CBMC) simulations[46−48] in the constant-volume Gibbs ensemble (GE)[49−51] and the isobaric−isothermal (NPT) ensemble[52] were carried out to compute the heat of vaporization and cohesive energy density, respectively. Table 1 provides a list of the 32 compounds studied in this work and of the version of the TraPPE force field and the system size used for the simulation of a specific compound. It should be noted that Belmares et al.[29] investigated a larger set of 64 compounds, but TraPPE parameters are not yet available for the remainder.

For the GEMC simulations, five different kinds of Monte Carlo moves were employed to sample the configurational part of the phase space: center-of-mass translations, rotations around the center of mass, conformational changes using CBMC,[46−48] CBMC particle swaps between the two boxes,[53,54] and volume exchanges between the boxes. For the special case of carboxylic acids that strongly associate in the vapor phase, aggregation-volume-bias MC moves[55,56] were also employed to sample the cluster distribution in the vapor phase. The maximum displacements for translational, rota-

Table 1. List of the 32 Compounds Studied, the Version of the TraPPE Force Field, and the System Size Used for the Monte Carlo Simulations

molecule	force field	system size
methanol	TraPPE-UA	450
ethanol	TraPPE-UA	400
1-propanol	TraPPE-UA	300
1-butanol	TraPPE-UA	250
butane-1,3-diol	TraPPE-UA	250
propane-1,2,3-triol (glycerol)	TraPPE-UA	250
1-pentanol	TraPPE-UA	250
2-ethylbutan-1-ol	TraPPE-UA	250
2-ethylhexan-1-ol	TraPPE-UA	200
diethyl ether	TraPPE-UA	250
n-hexane	TraPPE-EH	250
2,2-dimethylpropane	TraPPE-EH	250
acetonitrile	TraPPE-EH	400
propionitrile	TraPPE-EH	300
propanedinitrile (malononitrile)	TraPPE-EH	300
acetone	TraPPE-UA	400
butan-2-one	TraPPE-UA	250
4-methyl-2-pentanone	TraPPE-UA	250
2,6-dimethyl-4-heptanone	TraPPE-UA	250
N,N-diethylamine	TraPPE-EH	250
N,N-dipropylamine	TraPPE-EH	250
benzene	TraPPE-EH	250
toluene	TraPPE-UA	250
ethylbenzene	TraPPE-UA	250
chlorobenzene	TraPPE-EH	250
o-dichlorobenzene	TraPPE-EA	250
ethylchloride	TraPPE-UA	400
1-chlorobutane	TraPPE-UA	400
dichlorodifluoromethane	TraPPE-EH	400
carbontetrachloride	TraPPE-EH	400
ethanoic acid (acetic acid)	TraPPE-UA	400
propionic acid	TraPPE-UA	300

tional, and volume moves were adjusted to achieve 50% acceptance rates. To increase the sampling efficiency, different maximum displacements were used for translations and rotations in the vapor and liquid boxes. The probabilities for volume and swap moves were adjusted to give at least one volume move accepted every 10 MC cycles and one swap move accepted every 10−50 MC cycles.

Simulations were started by placing the molecules on a simple-cubic lattice, followed by 1000 MC cycles (where one cycle consists of *N*, the number of molecules, randomly selected trial attempts) at high temperature to melt the initial crystalline lattice. Five thousand MC cycles at a temperature close to the critical temperature were used to cool the system, followed by another 5000 MC cycles to reach the desired temperature. During these melting and cooling stages only translational, rotational, and conformational moves were used. Thereafter, the system was equilibrated at the desired temperature for at least 50 000 MC cycles using all five move types. During this period, the volume of the vapor box was adjusted to allow for an average of 20−40 molecules in the vapor phase. The production periods consisted of 50 000 MC cycles. The statistical uncertainties were determined by dividing the production run into 5 blocks.

The enthalpy of vaporization is computed on-the-fly in the GEMC simulation (after every MC move) using the formula

$$\Delta H_{\text{vap}} = \langle U_{\text{vap}} - U_{\text{liq}} + P_{\text{sat}}\Delta V\rangle_{\text{Gibbs}} \tag{5}$$

where U_{vap}, U_{liq}, P_{sat}, and ΔV are the instantaneous values of the molar internal energy of the vapor and liquid phases, the saturated vapor pressure, and the difference in the molar volume of the liquid and the vapor phase, respectively.

The isobaric−isothermal MC simulations employed translational, rotational, and conformational moves of single molecules and volume exchanges with an external pressure bath $P_{\text{ext}} = 1$ atm. The use of the Ewald summation to compute the electrostatic interactions (see below) makes it computationally expensive to separate the inter- and intramolecular components of the first-order electrostatic energy. Hence, an isolated molecule was simulated in a separate box to compute the average intramolecular energy. The solubility parameter at 1 atm was computed on-the-fly using the following equation

$$\delta_{\text{H}} = \left\langle \left(\frac{U_{\text{iso}} - U_{\text{liq}}}{V_{\text{liq}}} \right)^{1/2} \right\rangle_{\text{NPT}} \tag{6}$$

where U_{iso}, U_{liq}, and V_{liq} are the instantaneous values of the intramolecular energy (per mole) of the isolated molecule, the molar internal energy of the liquid phase, and the molar volume of the liquid phase, respectively.

For both GEMC and isobaric−isothermal simulations, a site−site based spherical potential cutoff at $r_{\text{cut}} = 14$ Å was used for the Lennard-Jones (LJ) interactions and the real space part of the Ewald summation. Analytical tail corrections[57] were used to account for the LJ potential beyond r_{cut}. The Ewald summation[51,57] with tin foil boundary condition was used to calculate the Coulombic interactions. The Ewald sum convergence parameter, κ, was set to $3.2/r_{\text{cut}}$, and the maximum number of reciprocal space vectors, $\mathbf{K}_{\text{max}}$, was set to 10.

In addition to the MC simulations for the TraPPE force field, we also carried out MD simulations following the protocol suggested by Belmares et al.[29] to compute the solubility parameters for the Dreiding force field with Mulliken or ESP partial charges at the normal boiling point. Complete details for this simulation protocol can be found in the original reference.

2.3. Force Field. The TraPPE force field[58] has been parametrized for a large set of organic compounds including linear and branched alkanes,[48,59,60] alkenes,[61] alcohols,[62] ethers,[63] aldehydes,[63] ketones,[63] acids,[64] esters,[65] amines,[66] amides,[66] nitroalkanes,[66] sulfides,[67] disulfides,[67] thiols,[67] and aromatic heterocycles.[68] The TraPPE force field derives its strength from the simplicity of the potential functions used and the transferability of interaction sites that allows for building of new compounds not included in the parametrization set. As for many other force fields, the development of the TraPPE force field involves fitting of the parameters for intermolecular interactions to experimental data. Whereas the development of prior force fields, such as the very successful optimized potentials for liquid simulations (OPLS) force field pioneered by Jorgensen and co-workers,[69] involves parametrization of the nonbonded parameters against experimental data (specific density, heat of vaporization, and heat capacity) of only the liquid phase at only one state point, the parametrization of the TraPPE force field involves phase equilibrium data at multiple state points. The use of vapor−liquid equilibrium properties became possible only with the emergence in the early 1990s of configurational-bias Monte Carlo simulations in the Gibbs ensemble[53,54] and quickly lead to a demonstration that force fields fitted at only one liquid-phase state point are often not as accurate over a more extensive region of the vapor−liquid coexistence curve.[70] This realization provided the incentive to fit nonbonded force field parameters against vapor−liquid coexistence curves.[71,72] Shortly after the first of these new force fields became available, it was also shown that they yield surprisingly accurate predictions of transport properties that are quite different from the equilibrium properties used for the parametrization.[42,43,73,74]

The development of the TraPPE force field involves a group-by-group parametrization philosophy that attempts to determine the parameters of a single group by fitting to the saturated liquid density, critical temperature, and saturated vapor pressure of a suitable test compound. The parameters for this group are then transferred when fitting the parameters for the next group, e.g., united-atom methyl group parameters fitted to the vapor−liquid equilibrium properties of ethane were used when fitting the parameters for the methylene group to the properties of *n*-pentane.[59] This group-by-group parametriztion philosophy yields in most cases unique parameters because only two Lennard-Jones parameters are fitted against a larger number of vapor−liquid equilibrium properties. However, single-component vapor−liquid equilibrium properties alone might not be sufficient for small polar molecules, e.g., carbon dioxide, ammonia, or benzene. In these cases, the parametrization either involved additional simulations for binary mixtures with alkanes[75] or for solid-fluid equilibrium properties.[76,77]

For alkanes, nitriles, and arenes, the TraPPE force field provides a choice of using either the united-atom (UA) or the explicit-hydrogen (EH) representation of CH_x groups. The UA version of the force field is simple and results in savings of computer time, while the EH hydrogen version provides more accurate vapor densities, vapor pressures, and heats of vaporization over a wide range of temperatures and pressures at a higher computational cost. In this work, we always use the EH version when available.

The total interaction energy for the TraPPE force field consists of bonded and nonbonded parts. The nonbonded interactions are represented by Lennard-Jones (LJ) and Coulomb potentials, given by

$$u(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{\mathrm{ij}}} \quad (7)$$

where r_{ij}, σ_{ij}, ϵ_{ij}, q_i, q_j, and ϵ_0 are the distance between two interaction sites, LJ size and LJ well depth, partial atomic charge on the interaction sites *i* and *j*, and the permittivity of the vacuum, respectively. The Lorentz−Berthelot combining rules[78] are used to determine LJ parameters for unlike interactions. In the TraPPE force field, molecules are treated as semiflexible chains: All bond lengths are kept rigid, bond bending angles are controlled by harmonic potentials, and dihedral motions are governed by cosine series potentials.

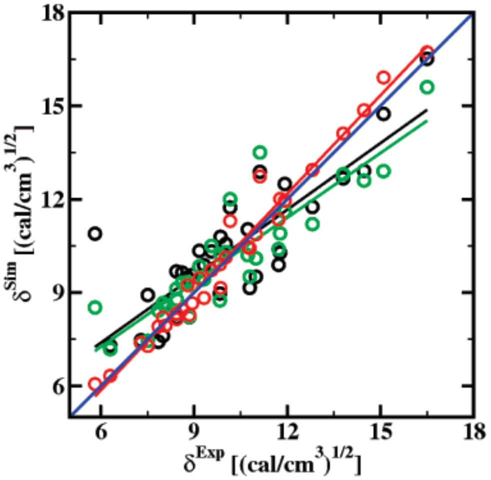


Figure 1. Comparison of the predicted solubility parameters at $T = 298.15$ K with experimental data.[29] The red, green, and black circles represent the solubility parameters computed with the TraPPE force field, the Dreiding force field with ESP charges, and the Dreiding force field with Mulliken charges. The correspondingly colored lines show the linear least-squares fits, and the blue line is the ideal correlation ($y = x$).

3. Results and Discussion

Figure 1 shows a comparison of the Hildebrand solubility parameters (numerical values are listed in Table 2) for the set of 32 compounds obtained from MC simulations using the TraPPE force field, MD simulations using the Dreiding force field with ESP (D/ESP) or Mulliken (D/MUL) partial charges,[29] the equation of state (EOS) approach,[23] and the experimental data.[29] This set includes a wide range of functional groups for which the TraPPE force field is available, namely alkanes, alcohols, ketones, ethers, nitriles, amines, benzene and benzene derivatives, alkylchlorides, fluoroalkanes, and carboxylic acids. The mean unsigned percent error (MUPE) in δ_H predicted using the TraPPE force field is 3.3%, whereas the errors for the D/ESP, D/MUL, and EOS are 9.2, 11.6, and 4.8%, respectively, i.e., the TraPPE force field performs significantly better.

A similar trend is observed from the correlation plot (see Figure 1). Linear least-squares fits yield $y = -0.453 + 1.05x$ (correlation coefficient, $R = 0.9855$), $y = 3.104 + 0.692x$ ($R = 0.897$), and $y = 3.113 + 0.713x$ ($R = 0.837$) for the TraPPE, D/ESP, and D/MUL force fields, respectively. The slope near unity, small intercept, and high R value obtained for the TraPPE force field shows that both absolute and relative δ_H values are predicted with excellent accuracy. In contrast, the D/ESP and D/MUL force fields yield slopes that are significantly smaller than unity and large positive intercepts, i.e., the relative accuracy is somewhat lacking. Nevertheless, the D/ESP force field performs somewhat better than D/MUL.

Table 2. Numerical Values of the Hildebrand Solubility Parameters (in Units of $(cal/cm^3)^{1/2}$) and Their Standard Deviations (SD)[a]

molecule	expt[b]	SD	TraPPE	SD	D/ESP[29]	SD	D/MUL[29]	SD	EOS[23]
methanol	14.5	0.08	14.9	0.05	12.6	0.71	12.9	0.55	14.6
ethanol	12.8	0.12	12.9	0.02	11.2	0.51	11.8	0.64	12.8
1-propanol	11.8	0.57	12.0	0.04	10.9	0.60	10.3	0.43	11.8
1-butanol	11.7	0.84	11.4	0.02	10.4	0.46	9.9	0.56	11.2
butane-1,3-diol[79]	13.8	6.16	14.1	0.09	12.8	0.52	12.7	0.33	
propane-1,2,3-triol	16.5	3.63	16.7	0.44	15.6	0.32	16.5	0.82	16.8
1-pentanol	11.0	0.99	10.9	0.02	10.1	0.48	9.5	0.18	10.7
2-ethylbutan-1-ol	10.8	0.84	10.4	0.05	9.5	0.35	9.2	0.46	
2-ethylhexan-1-ol	9.8	0.46	9.9	0.02	8.8	0.45	9.0	0.54	
diethyl ether	7.5	0.16	7.3	0.03	7.5	0.57	8.9	0.40	7.7
n-hexane	7.3	0.02	7.4	0.19	7.4	0.69	7.5	0.47	7.1
2,2-dimethylpropane	6.3		6.3	0.09	7.2	0.68	7.3	0.88	
acetonitrile	11.9	0.08	12.0	0.01	11.6	0.49	12.5	0.55	
propionitrile	10.7	0.07	10.5	0.01	10.2	0.46	11.0	0.33	
propanedinitrile	15.1		15.9	0.07	12.9	0.42	14.7	0.31	
acetone	9.8	0.16	9.1	0.02	10.2	0.59	10.8	0.50	9.8
butan-2-one	9.3	0.06	8.8	0.02	9.4	0.37	9.9	0.27	
4-methyl-2-pentanone	8.4	0.12	8.3	0.01	9.1	0.38	9.7	0.30	
2,6-dimethyl-4-heptanone	8.1	0.25	7.9	0.03	8.6	0.27	8.7	0.44	
N,N-diethylamine	8.0	0.03	8.2	0.02	8.7	0.58	7.6	0.33	8.2
N,N-dipropylamine	7.8	0.10	7.9	0.03	8.4	0.20	7.4	0.50	
benzene	9.2	0.04	9.5	0.02	9.8	0.51	10.3	0.47	8.9
toluene	8.9	0.08	8.7	0.02	9.2	0.66	9.6	0.20	8.7
ethylbenzene	8.8	0.04	9.2	0.00	9.3	0.41	9.5	0.26	
chlorobenzene	9.6	0.07	9.7	0.02	10.5	0.50	10.3	0.39	
o-dichlorobenzene	10.0	0.03	10.1	0.03	10.3	0.25	10.6	0.20	
ethylchloride	8.8	0.49	8.3	0.01	8.2	0.70	8.2	0.54	
1-chlorobutane	8.4		8.1	0.02	8.8	0.32	8.2	0.20	
dichlorodifluoromethane	5.8	0.44	6.1	0.02	8.5	0.40	10.9	0.36	
carbontetrachloride	8.6	0.05	8.5	0.02	9.3	0.29	9.6	0.45	
ethanoic acid	11.1	1.41	12.7	0.07	13.5	0.62	12.9	0.66	13.5
propionic acid	10.2	2.20	11.3	0.09	12.0	0.49	11.7	0.25	12.4
MUPE			3.3		9.2		11.6		4.8

[a] At $T = 298.15$ K and $P = 1$ atm. [b] Corrected average values taken from ref 29.

The accuracy of the EOS approach for δ_H predictions is close to that for the TraPPE force field. It should be noted that the EOS approach[23] uses characteristic parameters for each of the molecules, whereas the TraPPE force field has been fitted to specific functional groups, providing transferability of parameters. In the set of 32 molecules, EOS values were available only for 14 molecules, of which 6 were alcohols. When considering only the alcohols for which EOS values are available, the MUPEs for the TraPPE, D/ESP, D/MUL, and EOS are 1.9, 9.6, 10.1, and 1.7, respectively, i.e., the TraPPE force field and the EOS approach perform better for the alcohols than for the entire set.

None of the four approaches is able to predict δ_H within 5% of the average experimental values for the carboxylic acids. For the acids, the MUPEs for TraPPE, D/ESP, D/MUL, and EOS are 13, 20, 16, and 22%, respectively. While the EOS model performed extremely well for the alcohols, it is the worst for the acids. The TraPPE force field has the smallest MUPE. It should be noted that experimental values for acetic acid and propionic acid range from 10.1 to 13.0 and 8.1 to 12.7, respectively.[29] The large scatter in the experimental values is most likely caused by an inability to account for the extent of dimerization of smaller acids in the vapor phase. The TraPPE and D/MUL force fields predict δ_H for both acids within the experimental range. It is likely that these discrepancies will diminish for higher-molecular-weight carboxylic acids because their vapor pressures are sufficiently low that dimerization in the vapor phase is less prevalent.[64]

The numerical data listed in Table 3 and the correlation plots presented in Figure 2 clearly show that the force fields perform significantly better for predictions of the liquid density than of the solubility parameter. The MUPEs for the liquid densities obtained with the TraPPE, D/ESP, and D/MUL force fields are 1.3, 5.1, and 5.1, respectively, and the linear fits yield $y = 0.049 + 0.934x$ ($R = 0.997$), $y = -0.063 + 1.060x$ ($R = 0.957$), and $y = -0.091 + 1.096x$ ($R = 0.958$), respectively. In this case, all three force fields give slopes near unity and small absolute values for the intercept. However, there is substantially more scatter in the simulation data for the Dreiding force field as is evident from the lower correlation coefficients. For the prediction of liquid densities, there is no significant difference in the accuracy between D/ESP and D/MUL.

As the solubility parameters for most compounds are not available at higher temperatures, heats of vaporization at the

Table 3. Numerical Values of the Specific Densities (in Units of g/cm^3) and Their Standard Deviations (SD)[a]

molecule	expt[80]	TraPPE	SD	D/ESP[29]	D/MUL[29]
methanol	0.791	0.781	0.003	0.69	0.74
ethanol	0.794	0.780	0.002	0.72	0.76
1-propanol	0.804	0.797	0.002	0.71	0.72
1-butanol	0.810	0.803	0.001	0.75	0.75
butane-1,3-diol	1.005	1.002	0.002	0.91	0.94
propane-1,2,3-triol	1.261	1.174	0.003	1.09	1.14
1-pentanol	0.811	0.809	0.001	0.76	0.75
2-ethylbutan-1-ol	0.830	0.827	0.001	0.78	0.77
2-ethylhexan-1-ol	0.833	0.831	0.001	0.76	0.78
diethyl ether	0.706	0.706	0.001	0.69	0.75
n-hexane	0.659	0.653	0.001	0.65	0.65
2,2-dimethylpropane	0.580	0.608	0.004	0.64	0.63
acetonitrile	0.786	0.777	0.001	0.82	0.82
propionitrile	0.772	0.772	0.001	0.76	0.80
propanedinitrile	1.049	1.064	0.005	1.03	1.05
acetone	0.791	0.777	0.000	0.82	0.82
butan-2-one	0.805	0.789	0.001	0.80	0.81
4-methyl-2-pentanone	0.802	0.798	0.002	0.83	0.83
2,6-dimethyl-4-heptanone	0.808	0.806	0.002	0.82	0.81
N,N-diethylamine	0.707	0.702	0.001	0.71	0.65
N,N-dipropylamine	0.738	0.731	0.002	0.70	0.67
benzene	0.874	0.876	0.001	0.93	0.92
toluene	0.865	0.860	0.001	0.86	0.89
ethylbenzene	0.867	0.862	0.002	0.88	0.88
chlorobenzene	1.107	1.087	0.002	1.13	1.10
o-dichlorobenzene	1.306	1.278	0.002	1.31	1.30
ethylchloride	0.890	0.884	0.001	0.89	0.88
1-chlorobutane	0.886	0.872	0.001	0.87	0.86
dichlorodifluoromethane	1.310	1.296	0.002	1.56	1.62
carbontetrachloride	1.594	1.528	0.003	1.63	1.67
ethanoic acid	1.049	1.031	0.003	1.04	1.03
propionic acid	0.993	0.984	0.002	0.94	0.96
MUPE		1.3		5.1	5.1

[a] At $T = 298.15$ K and $P = 1$ atm.

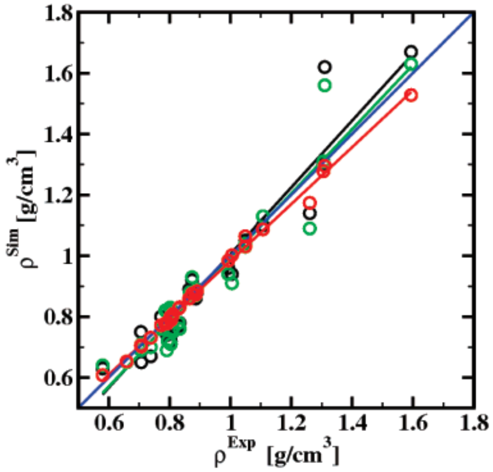


Figure 2. Comparison of the predicted specific densities at $T = 298.15$ K with experimental data.[80] The red, green, and black circles represent the specific densities computed with the TraPPE force field, the Dreiding force field with ESP charges, and the Dreiding force field with Mulliken charges. The correspondingly colored lines show the linear least-squares fits, and the blue line is the ideal correlation ($y = x$).

normal boiling point were used to test the efficacy of the force fields at higher temperatures. Table 4 lists the numerical data for ΔH_{vap} at 298.15 K and at the normal boiling point for each of the 32 molecules. It should be noted here that additional MD simulations were carried in this work to obtain the solubility parameters for D/ESP and D/MUL at the normal boiling point using the protocol of Belmares et al.[29] The values for δ_H were then converted to ΔH_{vap} using eqs 1, 3, and 4. The MUPEs for ΔH_{vap} predicted by TraPPE, D/ESP, and D/MUL at 298.15 K (and at T_b) are 4.5 (5.1), 11.2 (12.1), and 17.3 (19.8), respectively. Although the MUPEs give the impression that the accuracy of all three force fields does not deteriorate substantially when the temperature is increased, the correlation plots shown in Figure 3 indicate that this is not the case. The resulting least-squares fits for TraPPE, D/ESP, and D/MUL at 298.15 K are $y = 0.106 + 0.965x$ ($R = 0.982$), $y = 2.198 + 0.777x$ ($R = 0.934$), and $y = 2.885 + 0.733x$ ($R = 0.880$), respectively, and at the normal boiling point $y = 0.019 + 1.018x$ ($R = 0.941$), $y = 2.445 + 0.767x$ ($R = 0.808$), and $y = 4.745 + 0.489x$ ($R = 0.556$), respectively. The TraPPE force field predicts the heats of vaporization at 298.15 K and at T_b quite well as is evident from the slopes and the intercepts, and, as shown recently, the TraPPE force field also predicts the pressure dependence of the solubility parameter correctly.[82] The correlations for the D/ESP are of

Table 4. Numerical Values of the Heats of Vaporization (in Units of kcal/mol), at $T = 298.15$ K and at the Experimental Normal Boiling Point

	$T = 298.15$ K				T_b			
molecule	expt[81]	TraPPE	D/ESP[a]	D/MUL[a]	expt[81]	TraPPE	D/ESP[a]	D/MUL[a]
methanol	9.0	9.4	8.0	7.8	8.4	8.7	7.4	7.4
ethanol	10.0	10.3	8.6	9.0	9.2	9.3	8.3	7.7
1-propanol	11.3	11.4	10.6	9.4	9.9	9.8	10.0	7.9
1-butanol	12.2	12.3	11.3	10.3	10.3	10.3	10.4	9.0
butane-1,3-diol	17.7	18.2	16.8	16.0	13.0	14.2	13.2	12.8
propane-1,2,3-triol	21.9	21.4	21.2	22.6	n/a	15.8	16.3	13.0
1-pentanol	13.4	13.4	12.4	11.2	10.6	10.5	10.4	9.6
2-ethylbutan-1-ol	14.3	14.0	12.4	11.7	n/a	10.5	10.1	10.1
2-ethylhexan-1-ol	n/a	15.6	13.7	14.0	n/a	10.9	10.7	9.6
diethyl ether	6.5	6.1	6.6	8.5	6.3	6.0	6.3	9.1
n-hexane	7.6	7.6	7.8	8.0	6.9	7.0	6.9	6.9
2,2-dimethylpropane	5.3	5.5	6.4	6.7	5.4	5.6	6.4	6.4
acetonitrile	8.0	8.0	7.3	8.4	7.1	7.3	7.2	8.0
propionitrile	8.6	8.3	8.1	9.0	7.6	7.4	7.3	8.7
propanedinitrile	16.3	15.9	11.3	14.3	n/a	13.2	9.2	13.4
acetone	7.5	6.8	8.0	8.8	7.0	6.4	7.5	8.0
butan-2-one	8.3	7.7	8.6	9.3	7.5	7.0	7.8	8.9
4-methyl-2-pentanone	9.7	9.3	10.6	11.9	8.2	8.0	9.1	11.1
2,6-dimethyl-4-heptanone	12.2	11.7	13.3	13.8	n/a	9.4	10.2	11.4
N,N-diethylamine	7.5	7.7	8.3	7.1	6.9	7.1	7.4	6.6
N,N-dipropylamine	9.6	9.2	10.8	8.9	8.0	8.0	10.3	8.0
benzene	8.1	8.5	8.7	9.7	7.3	7.7	8.5	9.2
toluene	9.1	8.6	9.7	10.1	7.9	7.6	8.8	9.3
ethylbenzene	10.1	10.7	11.1	11.4	8.5	9.3	9.3	9.7
chlorobenzene	9.8	10.3	11.6	11.5	8.4	8.9	9.8	10.4
o-dichlorobenzene	n/a	12.4	12.5	13.2	n/a	10.2	11.6	12.1
ethylchloride	n/a	5.7	5.5	5.5	5.9	5.8	6.1	5.5
1-chlorobutane	8.0	7.7	8.8	7.8	7.3	7.0	7.9	7.1
dichlorodifluoromethane	4.1	4.0	6.2	9.4	4.9	4.8	6.5	8.7
carbontetrachloride	7.8	7.8	8.8	9.2	7.1	7.2	8.2	8.5
ethanoic acid	12.3	9.8	11.1	10.3	5.7	8.4	10.1	9.7
propionic acid	13.1	10.6	11.9	11.2	n/a	8.7	9.9	9.7
MUPE		4.5	11.2	17.3		5.1	12.1	19.8

[a] Computed from solubility parameters using eqs 1, 3, and 4.

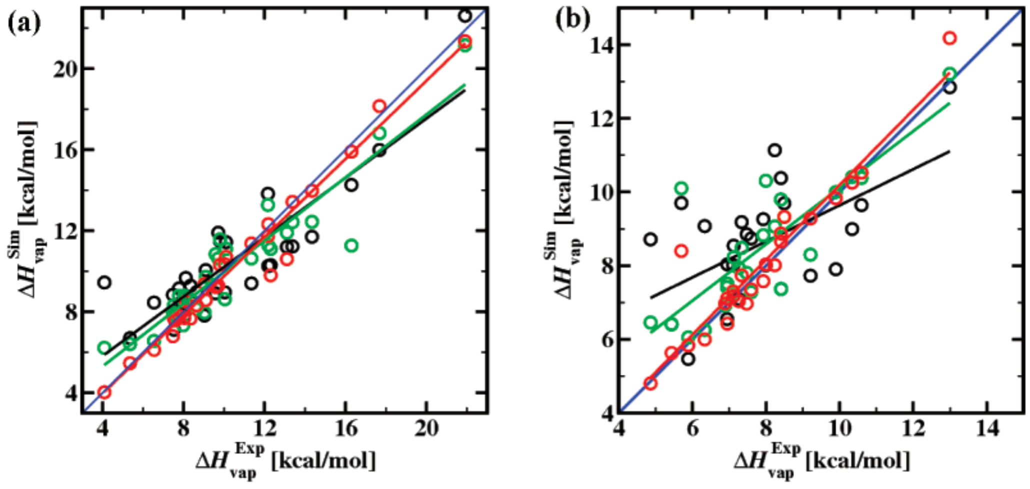


Figure 3. Comparison of the predicted heats of vaporization at $T = 298.15$ K (a) and the (experimental) normal boiling point (b) with experimental data.[81] The red, green, and black circles represent the specific densities computed with the TraPPE force field, the Dreiding force field with ESP charges, and the Dreiding force field with Mulliken charges. The correspondingly colored lines show the linear least-squares fits, and the blue line is the ideal correlation ($y = x$).

similar quality at both temperatures, the D/MUL force field performs significantly worse at elevated temperatures. It should be noted here that the deviations for the two acids are quite large. In particular, at T_b the TraPPE, D/ESP, and D/MUL force fields overestimate ΔH_{vap} by 50, 80, and 70%, respectively. Maybe, this points to a problem with the experimental data that show a decrease in ΔH_{vap} by an unusually large factor of 2 upon the increase in temperature.

4. Conclusions

The Hildebrand solubility parameters, liquid-phase densities, and heats of vaporization at the standard temperature, and the heats of vaporization at the normal boiling point were computed for a set of 32 common organic solvents and monomer units. The overall performance of the TraPPE force field is very satisfactory and significantly better compared to the Dreiding force field with either ESP or Mulliken partial charges. The main advantage of the Dreiding force field is that it can be applied to simulate a larger number of functional groups than are currently available for the TraPPE force field. The EOS approach is nearly as accurate as the TraPPE force field. The main drawback of the EOS approach is the lack of transferability due to the use of molecule-specific parameters. Based on the results presented here, we believe that molecular simulation offers a promising alternative to experimental measurements for the determination of solubility parameters for organic compounds.

Acknowledgment. Financial support from the National Science Foundation (CTS-0553911) and the 3M Company is gratefully acknowledged. Part of the computer resources were provided by the Minnesota Supercomputing Institute.

References

(1) Hildebrand, J. H.; Scott, R. L. *The Solubility of Nonelectrolytes*, 3rd ed.; Reinhold: New York, NY, 1950; pp 424−434.

(2) Zhu, S.; Paul, D. R. *Macromolecules* **2002**, *35*, 8227−8238.

(3) Mutelet, F.; Ekulu, G.; Solimando, R.; Rogalski, M. *Energy Fuels* **2004**, *18*, 667−673.

(4) Hancock, B. C.; York, P.; Rowe, R. C. *Int. J. Pharm.* **1997**, *148*, 1−21.

(5) Gu, C. H.; Li, H.; Gandhi, R. B.; Raghavan, K. *Int. J. Pharm.* **2004**, *283*, 117−125.

(6) Squillante, E.; Needham, T.; zia, H. *Int. J. Pharm.* **1997**, *159*, 171−180.

(7) Minghetti, P.; Cilurzo, F.; Casiraghi, A.; Montanary, L. *Int. J. Pharm.* **1999**, *190*, 91−101.

(8) Ray, S. K.; Sawant, S. B.; Joshi, J. B.; Pangarker, V. G. *Ind. Eng. Chem. Res.* **1997**, *36*, 5265−5276.

(9) Hirst, A. R.; Smith, D. K. *Langmuir* **2004**, *20*, 10851−10857.

(10) Lin, Y.; Alexandridis, P. *Langmuir* **2002**, *18*, 4220−4231.

(11) Jang, B. N.; Wang, D.; Wilkie, C. A. *Macromolecules* **2005**, *38*, 6533−6543.

(12) Bicerano, J. *Prediction of Polymer Properties*, 2nd ed.; Marcel Dekker: New York, NY, 1996; pp 108−136.

(13) Hansen, C. M. *J. Paint Technol.* **1967**, *39*, 104−117.

(14) Barton, A. F. M. *Chem. Rev.* **1975**, *75*, 731−753.

(15) Gardon, J. L. *J. Colloid Interface Sci.* **1977**, *59*, 582−596.

(16) Dipaola-Baranyi, G.; Guillet, J. E.; Klein, J.; Jeberien, H.-E. *J. Chromatogr.* **1978**, *166*, 349−356.

(17) Price, G. J.; Guillet, J. E. *J. Chromatogr.* **1986**, *369*, 273−280.

(18) Price, G. J.; Shillcock, I. M. *J. Chromatogr., A* **2002**, *964*, 199−204.

(19) Adamska, K.; Voelkel, A. *Int. J. Pharm.* **2005**, *304*, 11−17.

(20) Roberts, R. J.; Rowe, R. C. *Int. J. Pharm.* **1993**, *99*, 157−164.

(21) van Krevelen, D. W. *Properties of Polymers*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 1976; pp 129−153.

(22) Rogel, E. *Energy Fuels* **1997**, *11*, 920−925.

(23) Stefanis, E.; Tsivintzelis, I.; Panayiotou, C. *Fluid Phase Equil.* **2006**, *240*, 144−154.

(24) Theodorou, D. N.; Suter, U. W. *Macromolecules* **1985**, *18*, 1467−1478.

(25) Choi, P.; Kavassalis, T. A.; Rudin, A. *J. Colloid Interface Sci.* **1992**, *150*, 386−393.

(26) Lago, S.; Garzon, B.; Calero, S.; Vega, C. *J. Phys. Chem. B* **1997**, *101*, 6763−6771.

(27) Maranas, J. K.; Kumar, S. K.; Debenedetti, P. G.; Graessley, W. W.; Mondello, M.; Grest, G. S. *Macromolecules* **1998**, *31*, 6998−7002.

(28) Eichinger, B. E.; Rigby, D.; Stein, J. *Polymer* **2002**, *43*, 599−607.

(29) Belmares, M.; Blanco, M.; Goddard, W. A.; Ross, R. B.; Caldwell, G.; Chou, S. H.; Pham, J.; Olafson, P. M.; Thomas, C. *J. Comput. Chem.* **2004**, *25*, 1814−1826.

(30) Jawalkar, S. S.; Adoor, S. G.; Sairam, M.; Nadagouda, M. N.; Aminabhavi, T. M. *J. Phys. Chem. B* **2005**, *109*, 15611−15620.

(31) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(32) Prathab, B.; Aminabhavi, T. M.; Parthasarathi, R.; Manikandan, P.; Subramanian, V. *Polymer* **2006**, *1*, 1−11.

(33) Wick, C. D.; Siepmann, J. I.; Theodorou, D. N. *J. Am. Chem. Soc.* **2006**, *127*, 12338−12342.

(34) Zhang, L.; Siepmann, J. I. *J. Phys. Chem. B* **2005**, *109*, 2911−2919.

(35) Rai, N; Rafferty, J. L.; Maiti, A.; Siepmann, J. I. *Fluid Phase Equil.* **2007**, *260*, 199−211.

(36) Sun, L.; Siepmann, J. I.; Klotz, W. L.; Schure, M. R. *J. Chromatogr., A* **2006**, *1126*, 373−380.

(37) Zhang, L.; Rafferty, J. L.; Siepmann, J. I.; Chen, B.; Schure, M. R. *J. Chromatogr., A* **2006**, *1126*, 219−231.

(38) Sun, L.; Siepmann, J. I.; Schure, M. R. *J. Phys. Chem. B* **2006**, *110*, 10519−10525.

(39) Rafferty, J. L.; Zhang, L.; Siepmann, J. I.; Chen, B.; Schure, M. R. *Anal. Chem.* **2007**, *79*, 6551−6558.

(40) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **2006**, *110*, 3555−3563.

(41) Wick, C. D.; Siepmann, J. I.; Sheth, A. R.; Grant, D. J. W.; Karaborni, S. *Cryst. Growth Des.* **2006**, *6*, 1318−1323.

(42) Kioupis, L. I.; Maginn, E. J. *J. Phys. Chem. B* **1999**, *103*, 10781−10790.

(43) Kioupis, L. I.; Maginn, E. J. *J. Phys. Chem. B* **200**, *104*, 7774−7783.

(44) Moore, J. D.; Cui, S. T.; Cochran, H. D.; Cummings, P. T. *J. Chem. Phys.* **2000**, *113*, 8833−8840.

(45) Kelkar, M. S.; Rafferty, J. L.; Siepmann, J. I.; Maginn, E. J. *Fluid Phase Equil.* **2007**, *260*, 218−231.

(46) Siepmann, J. I.; Frenkel, D. *Mol. Phys.* **1992**, *75*, 59−70.

(47) Vlugt, T. J. H.; Martin, M. G.; Smit, B.; Siepmann, J. I.; Krishna, R. *Mol. Phys.* **1998**, *94*, 727−733.

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

(49) Panagiotopoulos, A. Z. *Mol. Phys.* **1987**, *61*, 813−826.

(50) Panagiotopoulos, A. Z.; Quirke, N.; Stapleton, M.; Tildesley, D. J. *Mol. Phys.* **1988**, *63*, 527−545.

(51) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, 2002; pp 201−224.

(52) McDonald, I. R. *Mol. Phys.* **1972**, *23*, 41−58.

(53) Mooij, G. C. A. M.; Frenkel, D.; Smit, B. *J. Phys.: Condens. Matter* **1992**, *4*, L255−L259.

(54) Laso, M.; Pablo, J. J. D.; Suter, U. W. *J. Chem. Phys.* **1992**, *97*, 2817−2819.

(55) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 8725−8734.

(56) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **2001**, *105*, 11275−11282.

(57) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U.K.; 1987; pp 64−65, 156−162.

(58) *Transferable Potentials for Phase Equilibria Force Field*. http://www.chem.umn.edu/groups/siepmann/trappe/intro.php (accessed June 1, 2007).

(59) Martin, M. G.; Siepmann, J. I. *J. Chem. Phys.* **1998**, *102*, 2569−2577.

(60) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **1999**, *103*, 5370−5379.

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

(62) Chen, B.; Potoff, J. J.; Siepmann, J. I. *J. Phys. Chem. B* **2001**, *105*, 3093−3104.

(63) Stubbs, J. M.; Potoff, J. J.; Siepmann, J. I. *J. Phys. Chem. B* **2004**, *108*, 17596−17605.

(64) Kamath, G.; Cao, F.; Potoff, J. J. *J. Phys. Chem. B* **2004**, *108*, 14130−14136.

(65) Kamath, G.; Robinson, J.; Potoff, J. J. *Fluid Phase Equil.* **2006**, *240*, 46−55.

(66) Wick, C. D.; Stubbs, J. M.; Rai, N.; Siepmann, J. I. *J. Phys. Chem. B* **2005**, *104*, 18974−18982.

(67) Lubna, N.; Kamath, G.; Potoff, J. J.; Rai, N.; Siepmann, J. I. *J. Phys. Chem. B* **2005**, *109*, 24100−24107.

(68) Rai, N.; Siepmann, J. I. *J. Phys. Chem. B* **2007**, *111*, 10790−10799.

(69) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638−6646.

(70) Siepmann, J. I.; Karaborni, S.; Smit, B. *J. Am. Chem. Soc.* **1993**, *115*, 6454−6455.

(71) Smit, B.; Karaborni, S.; Siepmann, J. I. *J. Chem. Phys.* **1995**, *102*, 2126−2140; **1998**, *109*, 352.

(72) Siepmann, J. I.; Karaborni, S.; Smit, B. *Nature* **1993**, *365*, 330−332.

(73) Mundy, C. J.; Siepmann, J. I.; Klein, M. L. *J. Chem. Phys.* **1995**, *103*, 10192−10200; **1998**, *104*, 7797.

(74) Mundy, C. J.; Klein, M. L.; Siepmann, J. I. *J. Phys. Chem.* **1996**, *100*, 16779−16781.

(75) Potoff, J. J.; Siepmann, J. I. *AIChE J.* **2001**, *47*, 1676−1682.

(76) Chen, B.; Siepmann, J. I.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 9840−9848.

(77) Zhao, X. S.; Chen, B.; Karaborni, S.; Siepmann, J. I. *J. Phys. Chem. B* **2005**, *109*, 5368−5374.

(78) Maitland, G. C.; Rigby, M.; Smith, E. B.; Wakeham, W. A. *Intermolecular Forces: Their Origin and Determination;* Oxford Science Publications: Oxford, U.K., 1987; p 519.

(79) Steele, W. V.; Chirico, R. D.; Knipmeyer, S. E.; Nguyen, A. *J. Chem. Eng. Data* **1996**, *41*, 1255−1268.

(80) *Aldrich Handbook of Fine Chemicals and Laboratory Equipment*; Sigma-Aldrich Co.

(81) Lemmon, E. W.; McLinden, M. O.; Friend, D. G. *Thermophysical Properties of Fluid Systems*; NIST Chemistry WebBook, NIST Standard Reference Database Number 69, Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD 20899, June 2005. http://webbook.nist.gov (accessed June 1, 2007).

(82) Rai, N.; Siepmann, J. I.; Schultz, N. E.; Ross, R. B. *J. Phys. Chem. C* **2007**, *111*, 15634−15641.

CT700135J

JCTC Journal of Chemical Theory and Computation

Computational Study of the Small Zr(IV) Polynuclear Species

Niny Rao,[†] Marian N. Holerca,[‡] and Vojislava Pophristic*,[†]

Department of Chemistry & Biochemistry and the West Center for Computational Chemistry and Drug Design, University of the Sciences in Philadelphia, Philadelphia, Pennsylvania 19104, and Colgate Palmolive Company, Global Technology Center, Piscataway, New Jersey 08854

Received May 9, 2007

Abstract: Despite widespread zirconium use ranging from nuclear technology to antiperspirants, important aspects of its solvation chemistry, such as the nature of small zirconium(IV) hydroxy cluster ions in aqueous solution, are not known due to the complexity of the zirconium aqueous chemistry. Using a combination of Car−Parrinello molecular dynamics simulations and conventional quantum mechanical calculations, we have determined the structural characteristics and analyzed the aqueous solution dynamics of the two smallest zirconium(IV) cluster species possible, i.e., the dimer and trimer. Our study points to and provides detailed geometrical information for a stable structural motif for building zirconium polymers, the $Zr(OH)_2Zr$ bridging unit with 7−8 coordinated Zr ions, which, however, cannot be used to construct a stable structure for the trimer. We find that a stacked trimer, not featuring this motif, is a possible structure, though not a very stable one, shedding new light on this species, and its possible importance in the aqueous chemistry of Zr^{4+} ion.

I. Introduction

The aqueous chemistry of metal cations is of great interest due to their important roles in chemistry, geochemistry, and biochemistry. The structures, charges, and stabilities of aqueous metal cations and their polynuclear hydrolysis products are crucial for understanding and controlling processes such as their adsorption onto soil/mineral particles; coagulation/precipitation; chemical separations; and interactions with living organisms. Understanding transition and inner transition metal hydrolysis presents a special experimental challenge, due to the complexity and variability of the species formed by these ions in water as well as radioactivity in some cases. Despite decades of research, triggered by applications ranging from drug design to nuclear technology, many physicochemical characteristics of these highly charged ions and their hydrolysis products remain unknown.

The principal experimental problems associated with studying the hydrolysis of highly charged cations are related to the variability and complexity of solution composition and the simultaneous presence of many diverse polynuclear hydrolysis products. Using computational methods, one can isolate a specific chemical species or a combination of species, control the system conditions, and make observations and analyses of processes at the atomic level.

We tackle here the characteristics of polynuclear species formed by solvation of a IVB group metal cation, Zr^{4+}. Our choice of the transition metal to study stems from the important uses of zirconium and the fact that its chemistry is representative of the other IVB group elements as well as our lack of understanding of certain aspects of Zr, Ti, and Hf (group IVB) chemical behavior.[1] For example, Zr^{4+}, Hf^{4+}, and Ti^{4+} polynuclear clusters have been recently found to bind to an Fe^{3+}-binding protein, a member of the transferrin superfamily,[2] which plays a role in biomineralization.[3,4] In addition, Zr^{4+} ion is an essential ingredient of all antiperspirants and thus interacts with human biochemistry through widespread and everyday antiperspirant use, although the

* Corresponding author phone: (215)596-8551; fax: (215)596-5432; e-mail: v.pophri@usip.edu.

[†] University of the Sciences in Philadelphia.

[‡] Colgate Palmolive Company.

10.1021/ct7001094 CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/13/2007

nature and the biochemical effect of the specific polynuclear species formed on our skin as a result of Zr^{4+} polymerization are not known.[5]

Most importantly, zirconium metal is a crucial part of zircaloy, the material used in the nuclear fuel rod cladding. Despite its exceptional corrosion resistance, there is an emerging need to model zirconium corrosion due to the long-term possibility that corrosion may lead to leaks of radioactive material, consequent environmental contamination, and ultimately exposure of individuals to radioactivity.[6] The rate of migration of heavy metal ions out of nuclear waste depositories, mining tailings, and coal mines is similarly dependent on the particular hydrolytic species that are present. Common to all the above problems is the need to know exactly what species form in aqueous solutions of these ions under the conditions present and what their physicochemical features are.

In solution, zirconium exists exclusively in a +4 state and is believed to attain coordination numbers of 7 and 8, higher than typical for 3d-transition elements.[7] As opposed to many other transition metals, due to the high charge/radius ratio, Zr^{4+} ion as well as the other two ions of the IVB group (Hf^{4+} and Ti^{4+}) strongly hydrolyzes in water, leading to the formation of polynuclear species with oxygen containing bridges.[7] With few exceptions, neither the structure nor the exact composition of the hydrolyzed mononuclear and polynuclear species have been established. The extent of polymerization depends on many experimental parameters (e.g., aging, temperature, pH, and concentration), resulting in species with very different compositions, often difficult or impossible to distinguish experimentally.[7,8]

Specifically, we present herein our Car–Parrinello molecular dynamics (CPMD)[9] study of the small Zr^{4+} polynuclear cluster species and their behavior in an aqueous environment. The method is uniquely suited to the problem of identifying and analyzing relatively small structures and their behavior in aqueous solutions. This is due to the fact that it does not employ empirically parametrized forces to govern atomic motion but rather determines them “on the fly”, along with the molecular dynamics (MD) simulation, from the electronic structure calculations. Thus, CPMD can yield conclusions about interactions between particles in solution as well as properly model important solution processes involving bond breaking and/or formation, such as water deprotonation, and polynuclear species formation and disintegration. The effectiveness of CPMD in ion solvation studies has been demonstrated for a number of cations, including Cu^{2+},[10] Na^{+},[11] Ca^{2+},[12,13] Mg^{2+},[14] Fe^{3+},[15] Y^{3+},[16] K^{+},[17] Al^{3+},[18] Li^{+},[19] and Be^{2+}.[20] CPMD has also been successfully used for identification of unknown structures[21] (including hydrolysis products of ions)[22,23] and studies of their characteristics.[24]

Herein we focus on the two smallest polynuclear species zirconium is thought to form upon dissolution in aqueous media: the dimer and the trimer. Despite the importance of zirconium, only the structure of the tetrameric species ($Zr_4(OH)_8(H_2O)_{16}Cl_8$) is known with certainty, from early X-ray[25,26] and other experiments.[27] X-ray scattering studies have shown that this species is the dominant form in solutions.[26,28]

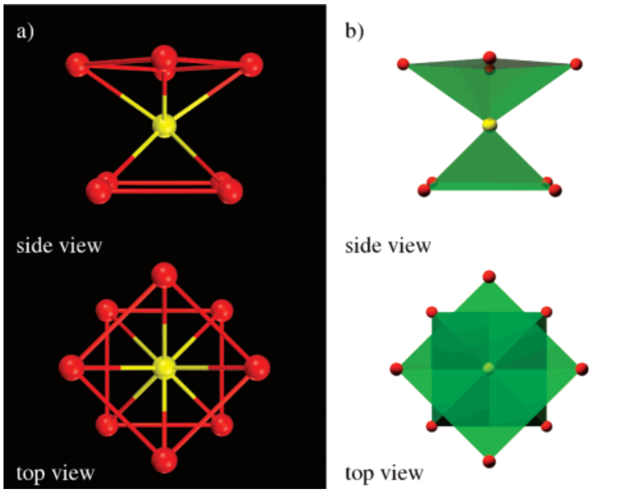


Figure 1. Square antiprism shape used as the building unit for the initial structure design of the dimer and trimer species (see also Figures 2, 3, 6, and 7). Two models (and two views) used further in the text are shown: (a) Ball and stick model - lines represent bonds between Zr^{4+} ion (yellow) and the bridging oxygen atoms (red) and H_2O moieties. (b) Planes cornered by Zr^{4+} ion (yellow) and oxygen atoms (red) represented in solid or translucent colors (green) to facilitate viewing of the structure in space, with perspective.

Spectrophotometric,[29] ultracentrifugation,[30] and light-scattering[31] studies suggest a possible presence of a trimer, $Zr_3(OH)_4^{8+}$, or $Zr_3(OH)_5^{7+}$. However, several recent publications find evidence for only the dimer ($Zr_2(OH)_6^{2+}$ and $Zr_2(OH)_7^{+}$) and tetramer.[1,32] We have determined and characterized the gas-phase and solution structures of the dimer and trimer species using the computational methods described herein and analyzed their stability and chemical behavior in aqueous solution.

II. Methodology

Computational studies of the zirconium system were performed using ab initio (Car–Parrinello) molecular dynamics (CPMD).[9] We employed a Goedecker-type pseudopotential for zirconium[33] and nonlocal norm-conserving soft pseudopotentials of Troullier-Martins type[34] for oxygen and hydrogen. Angular momentum components up to $l_{max} = 2$ have been included for Zr and $l_{max} = 1$ for O. The BLYP exchange correlation functional was employed,[35,36] along with a plane wave basis with a 70 Ry cutoff. All simulations were performed in a periodically repeating cubic box, with the size varying depending on the specific Zr system (see below), with periodic boundary conditions.

Initial structures of the two polymer classes were constructed using a square antiprism as the building unit (Figure 1), based on the expected 7–8 coordination of Zr^{4+} ion, the X-ray structure of the tetramer, and our CPMD simulations of the Zr^{4+} ion in solution.[37] In general, minimum energy geometries of gas-phase structures were obtained by an initial relaxation at 300 K (in some cases 100 and 50 K were used, see section III) for 4–5 ps, followed by simulated annealing and geometry optimization. Simulated annealing runs used scaling factors of 0.9998 and 0.9999 for ionic velocities (unless otherwise noted). The gas-phase simulation cell edge

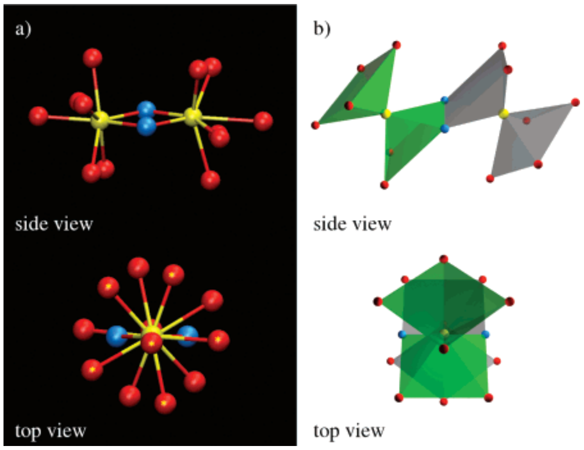


Figure 2. Initial structure of the staggered zirconium dimer. Two Zr^{4+} ions are connected by two O containing bridges and surrounded by additional six H_2O molecules each, in a staggered fashion with respect to the each other. (a) Ball and stick model. Yellow dots in the top view denote oxygen atoms bound to the front zirconium ion. (b) Two square antiprism monomer units are shown in different colors (green, gray) to facilitate viewing. Zr^{4+} ions: yellow; oxygen atoms in the bridging groups: blue; oxygen atoms in H_2O molecules: red. Side and top views are shown. Note that corners (bridge groups and H_2O molecules) of the antiprism units are staggered with respect to each other. Hydrogen atoms are not shown for clarity.

was 12.5 Å for the dimers, 14.0 Å for the compact trimers, and 18.0 Å for the linear trimers. In all calculations, classical equations of motion have been integrated with a velocity Verlet algorithm with a time step of 0.1207 fs and a fictitious mass for the electronic degrees of freedom of $\mu = 500$ au.

After optimization by the CPMD code, structures of the dimer and trimer species were refined by BLYP[38,39] and B3LYP optimizations,[39,40] using the LanL2DZ basis set.[41] These optimizations were carried out by Gaussian03.[42] Harmonic frequencies for the optimized geometries of these species have been calculated to ensure that they correspond to the local minima. The size of the polynuclear species prevented higher level optimizations.

Optimized gas-phase structures were used as the starting geometries for the simulations in aqueous solution. Such computations were undertaken for one dimer species (cubic box with a 12.5 Å edge and 49 H_2O molecules) and one form of the trimer (cubic box with a 15.6 Å edge and 77 H_2O molecules; for the exact form, see section III.B). Stabilities of these species in solution were determined by equilibrating the systems at 300 K using a Nosé-Hoover chain thermostat (of length 4, with frequency 500 cm^{-1})[43−46] for 10 ps.

III. Results

A. Dimer Clusters. The structure of the dimer, constructed as the starting point for the study, consists of two square antiprism units, with corners occupied by H_2O molecules and Zr ions in the centers of the units; these units are connected by two O containing bridges, with no direct Zr−Zr bond (Figure 2). Water molecules were chosen for the ligands following the structure of Zr^{4+} tetramer, which is the only Zr polynuclear species whose structure has been obtained both experimentally (e.g., X-ray studies) and by computational means ($Zr_2(OH)_6^{2+}$ and $Zr_2(OH)_7^{+}$, observed experimentally, form in solutions with pH<3.5). Several variants of the general dimer structure with respect to the bridge composition were subjected to the CPMD simulation in gas phase at 300 K. We tested the following structures: (a) a dimer with two OH bridges [$Zr_2(OH)_2(H_2O)_{12}^{6+}$]; (b) a dimer with an O and an OH bridge [$Zr_2O(OH)(H_2O)_{12}^{5+}$]; (c) a dimer with an H_2O and an OH bridge [$Zr_2(H_2O)(OH)(H_2O)_{12}^{7+}$]; and (d) a dimer with two O bridges [$Zr_2(O)_2(H_2O)_{12}^{4+}$]. The structure described under (c) proved to be unstable, as the H_2O bridge departed the cluster in the course of the gas-phase CPMD simulation (we observe the following reaction: $[Zr_2(H_2O)(OH)(H_2O)_{12}^{7+}] \rightarrow [(H_2O)_5Zr\text{-}(OH)\text{-}Zr\text{-}(H_2O)_5^{7+}] + 3H_2O$). In the other three cases, the dimer structures persist for about 3 ps (length of the simulation), in the staggered conformation as described below for the [$Zr_2(OH)_2(H_2O)_{12}^{6+}$] cluster. Due to the computational expense associated with CPMD simulations, we focused on only one structure for further investigations. The [$Zr_2(OH)_2(H_2O)_{12}^{6+}$] cluster was chosen based on its similarity to the Zr tetramer structure.[47]

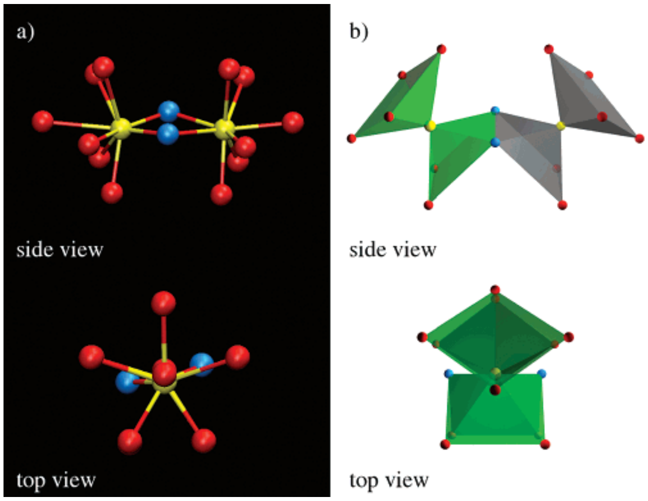


Figure 3. Initial structure of the eclipsed zirconium dimer. Two Zr^{4+} ions are connected by two O containing bridges and surrounded by additional six H_2O molecules each, in an eclipsed fashion with respect to the each other. (a) Ball and stick model. (b) Two square antiprism monomer units are shown in different colors (green, gray) to facilitate viewing. Zr^{4+} ions: yellow; oxygen atoms in the bridging groups: blue; oxygen atoms in H_2O molecules: red. Side and top views are shown. Note that the corners (bridging groups and H_2O molecules) of the antiprism units eclipse each other. Hydrogen atoms are not shown for clarity.

For the [$Zr_2(OH)_2(H_2O)_{12}^{6+}$] cluster, two initial conformations have been constructed (Figures 2 and 3). The first conformation, which will be referred to as the *staggered* conformation, has H_2O molecules which coordinate the two Zr^{4+} ions, 36° out of phase with each other (Figure 2). In the second conformation, the H_2O molecules surrounding two Zr^{4+} ions overlap each other (Figure 3), so this form is referred to as the *eclipsed* conformation. Both structures were

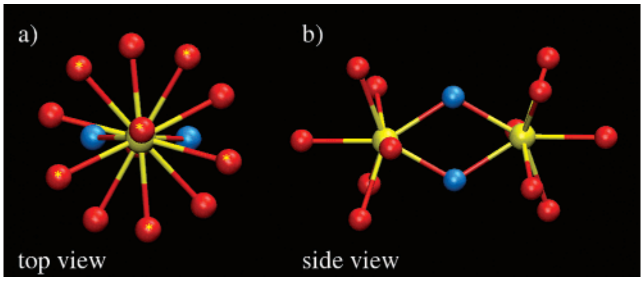


Figure 4. Optimized structure of the $[Zr_2(OH)_2(H_2O)_{12}{}^{6+}]$ dimer (BLYP/plane wave optimized structure shown; general features are the same as in the B3LYP/LanL2DZ and BLYP/LanL2DZ optimized ones; for the differences, see Table 1). Views shown: (a) view along the Zr−Zr axis. Only bonds connecting Zr^{4+} ions and oxygen atoms are shown. Yellow stars in the top view denote oxygen atoms bound to the front zirconium ion. (b) Side view, showing the plane consisting of two Zr^{4+} ions and two OH bridges. Zr^{4+} ions: yellow; oxygen atoms in OH bridging groups: blue; oxygen atoms in H_2O molecules: red.

subjected to relaxation at 300 K followed by simulated annealing in gas phase. During the relaxation phase (5 ps), the eclipsed conformation converted to the staggered conformation and remained stable throughout the simulated annealing. The staggered conformation, as expected, did not change its general features in the course of this procedure. Simulated annealing started from either the staggered or the eclipsed forms resulted in the same optimized structure, shown in Figure 4. These results indicate that the stable conformation of the dimer is the staggered one, as a consequence of a better accommodation of the steric crowding.

The annealed structure was optimized using plane wave basis, yielding a roughly symmetrical final conformation, with a Zr−Zr distance of 3.8 Å (Figure 4, Table 1). The two oxygen atoms in OH bridges are 2.2 Å apart from each of the zirconium ions. Oxygen atoms at the corners of the pyramid bases of the two units are staggered with respect to each other when viewed along the Zr−Zr axis. More precisely, oxygen atoms at the corners of the top monomer are 36° out of phase from those in the bottom monomer (Figure 4). The structure was also subjected to optimization using BLYP and the B3LYP/LanL2DZ level, with results in close agreement to the ones described above (Table 1). The obtained distance between Zr atoms and bridging O atoms is in the range of the values published for related Zr compounds in the solid state, such as the tetragonal (both experimental and calculated) ZrO_2 structure (2.09−2.44 Å)[48] and the calculated amorphous ZrO_2 (2.04−2.25 Å).[49] The Zr−Zr distance is, as expected, somewhat larger than in the monoclinic ZrO_2 structure (crystal structure data), which is 3.44−3.47 Å.[50] Our values also agree with the calculated gas-phase Zr−O bond lengths in a series of $Zr(OH)_n(H_2O)_m$ monomers with varying n, m (Zr−O_{OH}: 1.9−2.2 Å; Zr−O_{H2O}: 2.2−2.4 Å).[51]

In order to explore the dimer structure in aqueous solution, the optimized staggered structure was placed in a box with 49 H_2O molecules. The system was allowed to evolve for 10 ps at 300 K using a Nosé-Hoover chain thermostat. Analysis of the final trajectory revealed that both Zr−O radial distribution functions (for the two Zr^{4+} ions) have peaks at ∼2.2 Å, whereas the corresponding number of integrals show plateaus at 7 and 8, indicating different coordination numbers for the two Zr^{4+} ions (Figure 5a). Further examination of the trajectory revealed that one of the zirconium ions loses one of the initially eight coordinating water molecules within the first 1 ps of the simulation. We also note that the water molecules bound to the complex do not exchange with the bulk water molecules on the time scale of the simulation, as evident from the flat and long plateau of the radial distribution function (g(r)) at the zero value between the peaks for the first and the second shell of coordinated water (2.7−3.7 Å, Figure 5a). The Zr−Zr distance oscillates around 3.65 Å, with no drift (Figure 5b), indicating the general stability of the cluster. The first shell water molecules are organized around each zirconium ion in a pyramidal fashion (Figure 5c), the overall final arrangement being similar to that of the gas phase, except for one pyramid that lacks a water molecule. More specifically, the geometry of the first coordination shell of the eight-coordinated Zr ion corresponds to an antiprism, with the peaks of the angular distribution function (the plotted angle is ∠O−Zr−O) coinciding with those of the ideal antiprism (70°, 82°, 108°, and 142°). The peak at 70−85° (corresponding to the 70° and 82° peaks of the ideal antiprism) stems from two kinds of O−Zr−O angles: ∠O−Zr−O formed by the oxygen atoms which are next to each other and are located within the same pyramid (ideally 70°), and ∠O−Zr−O in which one oxygen atom is from the top and the other one is from bottom pyramid, and which are 45° out of phase from each other (ideally 82°). Another peak is centered at 140°, and is produced by the

Table 1. Geometrical Parameters of the Optimized Zr^{4+} Dimer ($Zr_2(OH)_2(H_2O)_{12}{}^{6+}$), Obtained Using Different Computational Levels[a]

	BLYP/plane wave basis	BLYP/ LanL2DZ	B3LYP/ LanL2DZ
	Distance (Å)		
Zr−Zr	3.751	3.871	3.819
Zr−O_{OH}	2.195−2.218	2.251−2.255	2.226−2.227
Zr−O_{H2O}	2.268−2.356	2.270−2.363	2.249−2.334
	Angle (deg)		
O_{OH}−Zr−O_{OH}	63.39−63.48	61.55	61.94
Zr−O_{OH}−Zr	116.44−116.69	118.44−118.46	118.06
O_{OH}−Zr−O_{H2O} (1)[b]	75.80−91.07	77.27−94.57	77.17−94.53
O_{OH}−Zr−O_{H2O} (2)[c]	79.48−79.70	79.86−81.99	79.78−81.98
O_{H2O}−Zr−O_{H2O} (1)[d]	71.47−71.52	71.93−71.94	71.89−71.92
O_{H2O}−Zr−O_{H2O} (2)[e]	71.71−81.16	71.33−79.18	71.30−79.07
O_{H2O}−Zr−O_{H2O} (3)[f]	70.30−73.97	69.68−73.27	69.76−73.29

[a] The optimized Zr dimer has a staggered configuration, in which two antiprism units are joined along the edge defined by the two OH bridges. The pyramids that contain this OH−OH edge are defined as base pyramids; the other two pyramids are defined as top pyramids. [b] O_{OH}−Zr−O_{H2O} (1) refers to angles defined by hydroxyl O, Zr, and water O atoms in the top pyramids. [c] O_{OH}−Zr−O_{H2O} (2) refers to angles defined by hydroxyl O, Zr, and water O atoms in the base pyramids. [d] O_{H2O}−Zr−O_{H2O} (1) refers to angles defined by water O atom in the base pyramids, Zr, and the other water O atoms in the same base pyramid. [e] O_{H2O}−Zr−O_{H2O} (2) refers to angles defined by water O atoms in the base pyramids, Zr, and water O atoms in the top pyramids. [f] O_{H2O}−Zr−O_{H2O} (3) refers to angles defined by water O atoms in the top pyramids, Zr, and water O atoms in the same top pyramid.

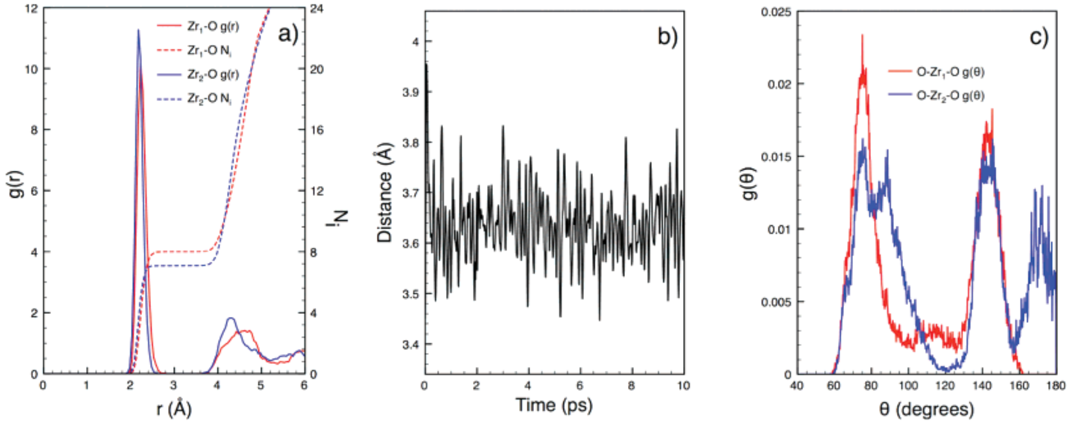


Figure 5. Zirconium dimer ($(Zr_2(OH)_2(H_2O)_{12}{}^{6+})$) in solution. (a) The Zr−O radial distribution functions (g(r), solid lines) and corresponding running coordination number integrals (NI, dashed lines) for the two Zr^{4+} ions. Note the difference between the coordination numbers for the two Zr^{4+} ions (7 and 8). (b) Zr−Zr distance as a function of time. (c) Angular distribution function. Angle plotted: ∠O−Zr−O. Arrangement of H_2O and OH groups around the eight-coordinated Zr ion (red) corresponds to an antiprism, whereas this arrangement around the seven-coordinated Zr ion (blue) corresponds to a pentagonal bipyramid.

O−Zr−O angle defined by oxygen atoms from the top and bottom pyramids, 135° out of phase from each other (ideally 142°). A much less pronounced feature at 105−115° comes from the O−Zr−O angle in which both oxygen atoms are located within the same pyramid and are opposite to each other (ideally 108°). The angular distribution function of the seven-coordinated Zr ion is clearly distinct from the eight-coordinated Zr case and indicates a pentagonal bipyramid geometry. Its coordinating groups produce peaks at 75° (angle between O atoms in the pentagonal base), 90° (angle between O atoms in pyramid plane and pyramid corner), 140° (angle between second nearest neighbors in the pentagonal base), and 180° (angle between two pyramid apexes), coinciding with the ideal pentagonal bipyramid peaks at 72°, 90°, 144°, and 180°. In both cases, departures from the ideal peak positions and widths are due to thermal fluctuations as well as the bending imposed by the existence of the two OH bridges.

Although data for direct comparison with experimental values are not available, the Zr−Zr and Zr−O distances fall within the range of values observed in dinuclear zirconium organometallic complexes, such as the one with the heptadentate ligand dhpta (reported average Zr−Zr and Zr−O distances are 3.5973 Å and 2.165 Å, respectively)[52] and the one with lactate ligands (3.5 Å and 2.0−2.2 Å, respectively).[53]

In summary, we find that a dimer structure with two OH bridges and 5−6 water molecules coordinating Zr^{4+} ions is stable both in gas-phase and aqueous solution, on the time scale of our simulation. The major difference between the gas-phase and aqueous solution results is in that the aqueous structure has one seven-coordinated and one eight-coordinated Zr^{4+} ion, as opposed to the two eight-coordinated Zr ions in gas phase. The arrangement of the terminal water molecules and OH groups within the monomer units is either square antiprism or pentagonal bipyramid, depending on the coordination, and the spatial relationship between the two units is such that water molecules are staggered.

B. Trimer Clusters. With an increasing number of Zr^{4+} ions, possibilities for different arrangements of the monomer units increase rapidly. Although more trimer configurations are conceivable, we focus here on two general configurations: linear and stacked (Figures 6 and 7). The linear structure consists of three antiprism units connected by bridges (Figure 6); the more compact, stacked structure consists of three monomer units joined by three bridges, each shared by two adjacent Zr^{4+} ions, and one, central bridge which is shared by all three Zr^{4+} ions (Figure 7). We attempted several compact trimer structures with respect to the nature of the bridges (the H_2O bridges were not considered based on the dimer result described in the previous section): (a) all four bridges are O^{2-} ions; (b) all four bridges are OH groups; (c) the central bridge, shared by three Zr^{4+}

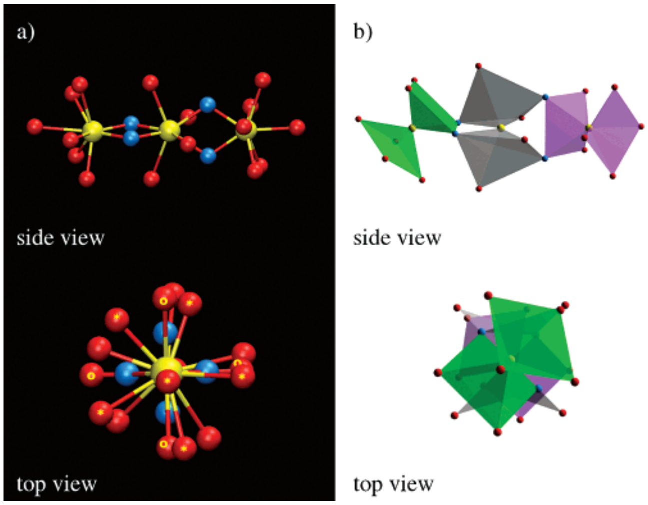


Figure 6. Initial structure of the linear trimer ($[Zr_3(OH)_4(H_2O)_{16}]^{8+}$). Three Zr^{4+} ions are connected by four OH bridges, in a linear fashion. (a) Ball and stick model. Yellow stars in the top view denote oxygen atoms bound to the front zirconium ion, whereas open yellow circles denote oxygen atoms bound to the middle Zr^{4+} ion. (b) Three square antiprism monomer units are shown in different colors (green, gray, purple) to facilitate viewing. Zr^{4+} ions: yellow; oxygen atoms in OH bridging groups: blue; oxygen atoms in H_2O molecules: red. Side and top views are shown.

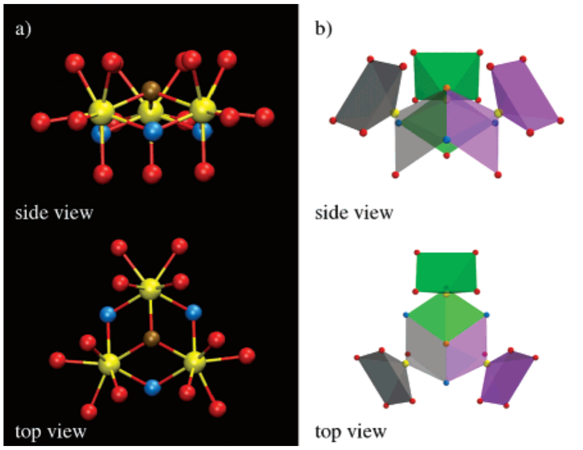


Figure 7. Initial structure of the stacked trimer. Three Zr^{4+} ions are connected by three bridges, and share another bridging group through a second set of three bridges. (a) Ball and stick model. (b) Three square antiprism monomer units are shown in different colors (green, gray, purple) to facilitate viewing. Zr^{4+} ions: yellow; oxygen atoms in bridging groups shared by two Zr^{4+} ions each: blue; oxygen atom bound to all three Zr^{4+} ions: brown; oxygen atoms in H_2O molecules: red. Side and top views are shown.

ions, is an OH group (brown, Figure 7); the three bridges, shared by two Zr^{4+} ions each (blue, Figure 7), are O^{2-} ions; (d) the central bridge is an O^{2-} ion, and two out of three bridges shared by two Zr^{4+} ions each are OH groups, with the remaining bridge an O^{2-} ion; (e) the central bridge is an O^{2-} ion, and two out of three bridges shared by two Zr^{4+} ions each are O^{2-} ions, with the remaining bridge an OH group; and (f) the central bridge is an O^{2-} ion, and the three bridges shared by two Zr^{4+} ions each are OH groups.

The initial configuration of the linear trimer was constructed by placing three Zr monomer units next to each other so that the two sets of OH bridges (connecting the first and the second unit, and connecting the second and the third unit) are 90° out of phase with respect to each other to minimize steric crowding effects. The Zr−Zr−Zr angle was set to 180°. Each Zr^{4+} ion was surrounded by water molecules (six and four for the terminal and middle Zr^{4+} ions, respectively) so that the coordination of eight was achieved. In addition, a bent linear structure was studied in which the Zr−Zr−Zr angle was set to 132°.

The stacked trimer was constructed by arranging each of the monomer units around three adjacent faces of a cube. The joint vertex of these three faces of the cube is occupied by the bridging group shared by three Zr^{4+} ions, while the bridges that connect two Zr^{4+} ions each occupy the remaining vertices (Figure 7). The Zr ions are placed above the center of each face, and additional water molecules are added so that each monomer unit is eight coordinated in an antiprism configuration. The zirconium ions are placed at an equal distance of 3.76 Å from each other. Intuitively, the stacked structure (shown in Figure 7) should be more stable than the linear one, due to its more compact geometry and additional linking bridges with respect to the linear trimer (6 bridges in the stacked trimer vs 4 bridges in the linear one). Such a stability order was also observed in a computational study of Al^{3+} cluster species.[54]

We first attempted to obtain a stable trimer structure in a fashion similar to the one used for the dimer: the described initial configurations were allowed to relax at 300 K in the gas phase. In the course of the gas-phase CPMD simulation at 300 K, with the exception of the above-described structure (b), which completely falls apart, all attempted stacked trimers follow a similar behavior, in that two Zr^{4+} ions remain at the distance of 3.0−3.2 Å, throughout the 7 ps simulation time, whereas one Zr^{4+} ion moves to a longer distances of 3.6−4.0 Å from the other two ions, extending the related bridges. Due to the computational cost of CPMD simulations (especially in aqueous solutions), we have chosen one of the above-described structures ($[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$) to extensively investigate with respect to both the gas-phase and aqueous solution dynamics (the choice is based on the structural relationship to the Zr^{4+} hexamer[55]). As we will discuss later in the text, on the CPMD accessible time scale, the trimer species exhibit a "breathing" behavior, in which the Zr^{4+} ion which initially moves somewhat away from the other two Zr^{4+} ions periodically comes back forming roughly the initial stacked trimer structure.

In an attempt to further examine the stability of the stacked trimer with respect to the Zr^{4+} ion drifting away from the cluster, the simulation for $[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$ was repeated at 100 K and at 50 K, neither of which was able to capture the anticipated degree of stability (i.e., no drifting of the Zr^{4+} ion). We also used the following procedure: all of the distances between Zr ions and O atoms were constrained to 2.3 Å, whereas O−H bonds as well as the corresponding bond angles and dihedral angles were permitted to change in the course of the simulation. Thus the water molecules around each Zr ion were allowed to relax to their optimal positions, while the monomer units were forced to remain bound to each other. The constrained CPMD simulation was carried out for 5 ps at 300 K and was followed by unconstrained simulated annealing using a scaling factor of 0.999. The stacked trimer remained mostly intact with only a H_3O^+ moiety leaving the main cluster. The geometrical/structural changes that occurred during the gas-phase simulation are described below.

With respect to the linear trimer, a "straight" and a "bent" form were tested. Both forms fall apart in the initial stages of CPMD gas-phase simulation. We attempted the constrained CPMD simulation as described for the stacked trimer, but, in the course of simulated annealing, the trimers disintegrated by the cleavage into a monomer and a dimer occurring at the OH bridges (resulting in one OH group leaving with the monomer unit and the other one remaining with the dimer). The geometry of the dimer is similar to the one described in section III.A. In conclusion, no stable linear trimer was observed.

The stacked trimer obtained from the constrained CPMD followed by simulated annealing had the following properties. The final configuration retained roughly its equilateral triangular shape, with Zr−Zr distances of 3.60, 3.54, and 3.51 Å. However, two of the zirconium ions have the coordination number of 7, while the third zirconium ion has

Table 2. Optimized Geometry of the $[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$ Trimer, Obtained Using Different Computational Levels[a]

	BLYP/plane wave basis	BLYP/ LanL2DZ	B3LYP/ LanL2DZ
	Distance (Å)		
Zr−Zr	3.487−3.604	3.588−3.607	3.553−3.572
$Zr-O_{OH}$	1.903−2.221	1.914−2.318	1.901−2.296
$Zr-O_{bridge}$	2.074−2.291	2.065−2.193	2.149−2.167
$Zr-O_{H2O}$	2.193−2.368	2.241−2.356	2.224−2.327
	Angle (deg)		
Zr−Zr−Zr	57.97−61.20	59.75−60.26	59.74−60.28
$O_{OH}-Zr-O_{OH}$	106.56−111.67	100.18−113.94	106.90−108.64
$Zr-O^{2-}-Zr$	110.54−114.33	111.38−115.67	111.70−115.72
O−Zr−O[b]	65.98−85.14	65.28−91.29	65.42−90.94
O−Zr−O[c]	106.29−162.91	100.18−172.05	100.86−171.50
O−Zr−O[d]	74.63−133.83	71.24−134.17	71.61−134.26
O−Zr−O[e]	146.59−152.79	147.05−151.63	147.12−151.86

[a] See text for the discussion about differences between the optimized structures. [b] O−Zr−O refers to angles defined by O, Zr, and O atoms in which the Zr and O atoms lay on the pentagonal base, and O atoms are adjacent to each other. The value for such an angle in an ideal pentagonal bipyramid is 72°. [c] O−Zr−O refers to angles defined by O, Zr, and O atoms in which the Zr and O atoms lay on the pentagonal base, and O atoms are not adjacent to each other. The value for such an angle in an ideal pentagonal bipyramid is 144°. [d] O−Zr−O refers to angles defined by an O atom located on the pentagonal base, and Zr and O atoms are located below or above the pentagonal base. The value for such an angle in an ideal pentagonal bipyramid is 90°. [e] O−Zr−O refers to angles defined by an O atom located below the pentagonal base, a Zr atom within the base, and an O atom located above the base. The value for such an angle in an ideal pentagonal bipyramid is 180°.

a final coordination number of 8. During the simulated annealing, one of the water molecules initially surrounding a Zr ion moved away from the first coordination shell, settling at the distance of 10.64 Å from the Zr^{4+} ion it originated from, while abducting a proton from a water molecule remaining within the cluster (bound to the same Zr ion). Subsequently, a second water molecule broke away from another Zr^{4+} ion and reached a stable distance of 4.5 Å from the Zr^{4+} ion it was bound to and that same distance to the eight-coordinated Zr^{4+} ion. Thus, the resulting cluster has two seven-coordinated Zr ions, one of which has a terminal OH group instead of a H_2O molecule.

The annealed structure was subjected to optimization at the B3LYP/LanL2DZ level, with the resulting geometrical parameters summarized in Table 2 (also see Figure 8). During the B3LYP optimization, one H_3O^+ moiety moved away from the cluster, resulting in a stacked trimer with three seven-coordinated Zr units. In other words, in addition to the two OH^- and one O^{2-} bridges, two of the three Zr ions have one OH group and three water molecules in the coordination shell. This structure was confirmed by BLYP/LanL2DZ optimization (Table 2). The BLYP/plane wave optimization was conducted starting from the geometry taken at the 10th ps of the CPMD simulation following the simulated annealing (discussed in the next paragraph). The difference with respect to the B3LYP/LanL2DZ geometry is that only one Zr ion has a coordination shell with an OH group. Other geometrical parameters are in agreement with the two other reported levels (Table 2). All Zr ions in the three optimized structures have a distorted pentagonal bipyramid configuration, with a range of values for the angles, centered at the value of the ideal pentagonal bipyramid (see footnote, Table 2). Further comparison with literature data is not possible, since the literature information is scarce and conflicting: no structure description has been published, and several compositions have been suggested and disputed.[1,56,57]

To determine the stability of the observed structure, the stacked trimer obtained from the simulated annealing process described above was simulated further. After a 5 ps gas-phase equilibration at 300 K using a Nosé-Hoover chain thermostat, the system was allowed to evolve for 50 ps. During the 50 ps CPMD simulation, a third water molecule (in addition to the two water molecules which left the cluster during simulated annealing) moved away from the first coordination shell of the remaining eight-coordinated Zr ion, to a distance of 4.15 Å from the nearest Zr^{4+} ion, making all three Zr^{4+} ions seven-coordinated, the same as that obtained by geometry optimizations (Figure 9a). The remaining bound water molecules produce a sharp g(r) peak at 2.2 Å distance for each of Zr^{4+} ions, with the running coordination number plateaus clearly at seven. In the case of the Zr^{4+} ion which has a terminal OH group in its first coordination shell, we observe two peaks in the radial distribution function: one at 1.8 Å, corresponding to the oxygen atom of the terminal (not a bridging one) OH group, and another one at 2.2 Å, stemming from oxygen atoms of terminal water molecules. Also, after the initial 11 ps of the simulation, the Zr^{4+} ion unit with one terminal OH^- group started drifting away from the cluster. It settled at a distance of ∼4.1 Å from the other two Zr^{4+} ions. This configuration persisted for about 6 ps, when the Zr^{4+} ion unit drifted back to ∼3.6 Å from the other two Zr^{4+} ions, and remained in this configuration for about 2 ps, followed by another movement to ∼4.1 Å from the other Zr^{4+} ions, where it again remained for ∼5 ps. After coming close to the rest of the cluster (remaining there for ∼4 ps), this unit departed the cluster again, this time remaining at ∼4.1 Å for ∼15 ps (Figure 9b), afterward joining the cluster for >6 ps. Thus, we observe an irregular oscillatory movement of one Zr unit with respect to the other two. At each instance of this Zr^{4+} ion moving away, the two other Zr^{4+} ions come closer to each other with respect to their distance in the initial several picoseconds of the simulation, as a consequence of decreased steric crowding. Moreover, the central bridging oxygen atom was observed to move above and below the plane defined by the three Zr^{4+} ions (Figure 9c) during the first 10 ps of simulation. From this point onward, this oxygen atom was found to remain on one side of the plane defined by the three Zr^{4+} ions. Thus, at all the times, one atom or a group of atoms was moving away from the rest of the cluster: in the first 10 ps, it was the central, bridging O atom; afterward, it was a Zr monomer unit.

The stacked trimer structure was then tested in solution, by placing the structure obtained at the end of the gas-phase constrained dynamics run in a box with 73 water molecules (15.6 Å size) and simulating the system for 10 ps. We find that all three Zr ions remain seven-coordinated (Figure 10a), with oxygen atoms of the coordinating species at 2.2 Å from

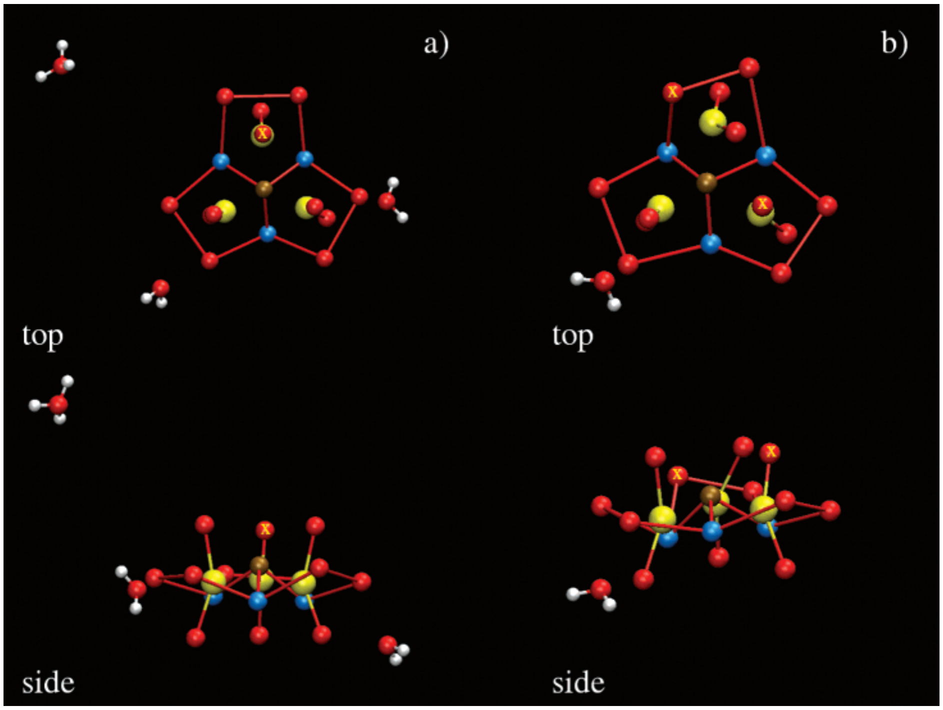


Figure 8. Optimized structure of the stacked trimer ($[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$). (a) Structure optimized by BLYP/plane wave basis set. (b) Structure optimized using the B3LYP/LanL2DZ basis set. The connecting lines show the pentagonal base in the pentagonal bipyramid arrangement of each Zr unit (not the chemical bonds). The Zr^{4+} ions are located at the center of each base. The axial oxygen atoms are connected to the Zr ions to show their axial position with respect to the pentagonal base. Zr^{4+} ions: yellow; oxygen atoms in OH bridging groups: blue; oxygen atom bound to all three Zr^{4+} ions: brown; oxygen atoms in H_2O molecules: red. The oxygen atoms in terminal OH groups are marked by a yellow "x" sign. Side and top views are shown.

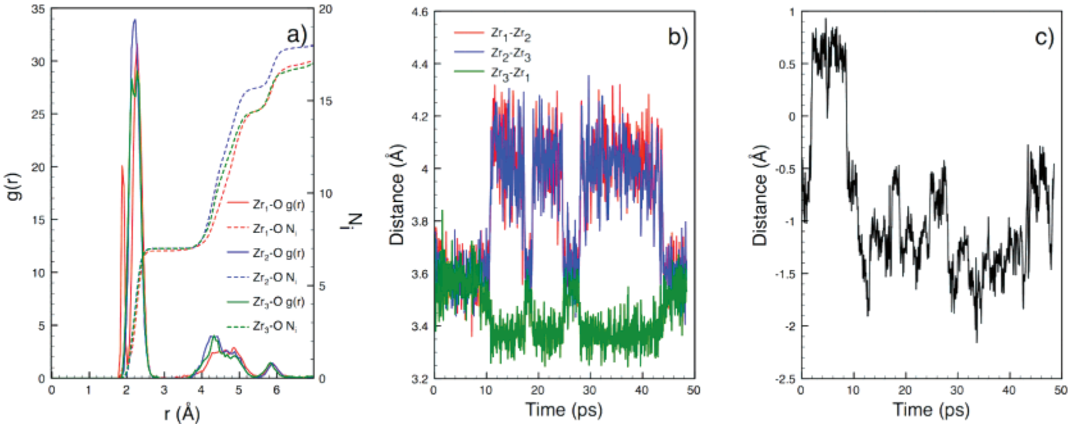


Figure 9. CPMD simulation of the stacked trimer ($[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$) in gas phase. (a) Zr−O radial distribution function (g(r), solid line) and the corresponding running integral coordination numbers (NI, dashed line) for the three Zr^{4+} ions. (b) Zr−Zr distance for all three possible pairs of Zr^{4+} ions (red, blue, and green lines). Note that one of the Zr^{4+} ions (red, blue) moves ~4.2 Å away from the other two in the 11th, 19th, and 28th ps of the simulation and remains at that distance for ~7 and ~16 ps. As this happens, the distance between the other two Zr^{4+} ions shrinks slightly, as a consequence of decreased steric crowding. (c) Distance of the O atom shared by the three Zr ions from the plane defined by the three Zr ions. The change of sign from positive to negative indicates that the O atom is moving from one side of the plane to the other.

the Zr ion. On the scale of the simulation, no exchange between the terminal and bulk water molecules is observed. Figure 10b illustrates the stability of the cluster within the 10 ps simulation time—the Zr−Zr distances oscillate around 3.54 Å, with no drift. The spatial arrangement of the coordinating species (one O bridge, two OH bridges, and four H_2O molecules) of the three seven-coordinated Zr ions is somewhat different from the dimer seven-coordinated unit: the locations of the peaks of the angular distribution functions point to a significantly distorted pentagonal bipyramid (Figure 10d). The ~72° peak, characteristic of the pentagonal bipyramid (angle between the species on the pentagonal base) is present. The main distortion from the ideal pentagonal bipyramid occurs with the position of the oxygen atoms perpendicular to the pentagonal base (axial oxygen atoms); instead of a 90° angle with the base, we

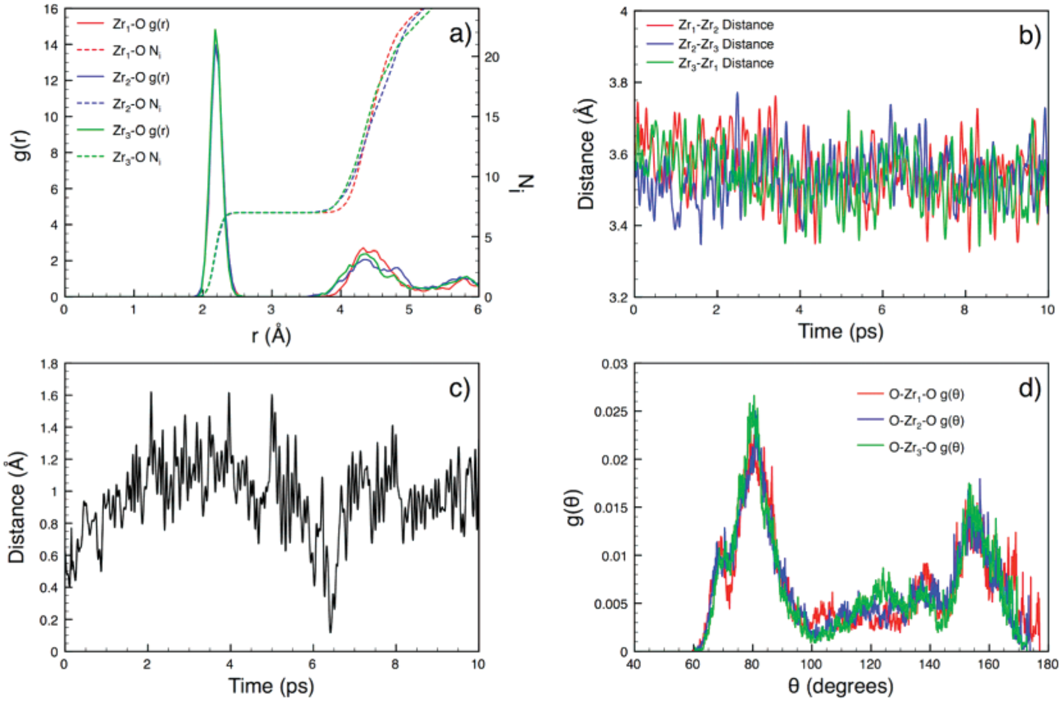


Figure 10. CPMD simulation of the stacked trimer ($[Zr_3(OH)_3O(H_2O)_{18}]^{7+}$) in aqueous solution. (a) The Zr−O radial distribution functions (g(r), solid lines) and corresponding running coordination number integrals (NI, dashed lines) for the three Zr^{4+} ions, indicating their seven coordination. (b) Zr−Zr distance for the three pairs of the Zr ions, oscillating around 3.54 Å. (c) Distance of the O atom shared by the three Zr ions from the plane defined by the three Zr ions. Unlike the gas-phase simulation, the O atom did not move from one side of the plane to the other. (d) Angular distribution function for the three Zr ions (O−Zr−O angle plotted), indicating a distorted pentagonal bipyramid arrangement of the coordinating oxygen atoms around each of the Zr ions.

observe an 85° angle. In addition, instead of an ideal linear arrangement (180°) between the two axial oxygen atoms around the same Zr ion, we find a peak at 155°.

In summary, we find that the stacked trimer is a possible, but not very stable, structure in the gas phase. A notable dynamical feature of the gas-phase structure is the oscillation of one monomer unit (shown here through the motion of the corresponding Zr^{4+} ion, Figure 9b), which might ultimately lead to cluster disintegration on a longer time scale. However, CPMD simulations beyond ∼50 ps presented here are not practical at this point in time. This motion was not observed in the course of the simulation in aqueous solution, due to the shorter simulation time as well as constraints posed by water molecules that surround the cluster (longer simulation, impractical at this time, is predicted to show one monomer unit leaving the cluster as well). In the course of the 10 ps simulation in solution, the species appears stable, with Zr−Zr distances oscillating around 3.54 Å. All three Zr ions remain surrounded by seven ligands (a bridging O group, two bridging OH groups, and four H_2O molecules; unlike the gas-phase structure, no terminal OH groups were found in the first coordination shell), arranged in a distorted pentagonal bipyramid geometries.

IV. Discussion

The present CPMD simulations, along with those presented in ref 47 show that several small Zr^{4+} polynuclear clusters exist in aqueous solution as structures consisting of seven- and eight-coordinated monomer units, with common features. Whereas the dimer and the tetramer[47] appear stable at the picosecond time scale, the trimer undergoes internal motions which indicate a possible cluster disintegration at a later stage.

The basic structural motif seen in the studied forms is a seven to eight coordinated Zr^{4+} ion, consistent with the coordination assigned to the Zr^{4+} ion in general.[7] The zirconium ion is surrounded by H_2O molecules and OH (or O^{2-}) groups, with Zr−O bonds of ∼2.2 Å length. After a certain coordination is assumed in the course of gas-phase structure determination, the initial changes in the coordination shells and settling to a certain coordination number, we do not observe Zr−O bond breaking, i.e., exchange of the terminal, bound H_2O molecules with the bulk. Depending on the coordination, the monomer units take either a pentagonal bipyramid (seven-coordinated) or antiprism spatial arrangements (eight-coordinated), the latter also observed in the case of the tetramer.[47] The strain imposed by the binding pattern between the units induces different degrees of distortion in these geometries.

The monomers are bound by O-containing bridges, resulting in another repeating, and stable structural motif. For the dimer, we focus on the $Zr(OH)_2Zr$ unit, which also appears in the tetramer species. However, such a pattern does not hold the molecule together in the case of the studied linear trimer: the $Zr(OH)_2Zr(OH)_2Zr$ "backbone" of the trimer breaks into a dimer and a monomer unit within a very short time. A structure that seems to better accommodate the steric crowding present in the trimer involves a bridging O atom, which connects to all three Zr ions, and single OH or O bridges between each pair of Zr^{4+} ions. The $Zr(OH)_2Zr$

structural motif has a consistent geometry when the dimer and the tetramer are compared: the Zr−Zr distance is ∼3.8 Å and the $Zr-O_{OH}$ distance is ∼2.2 Å. These distances are shorter and more dispersed in the thoroughly studied stacked trimer due to a much higher strain (3.5−3.6 Å Zr−Zr distance and 1.9−2.2 Å $Zr-O_{OH}$ distance). Such a highly strained trimer geometry is the probable cause of the fluctuations of one of the monomer units.

As opposed to the tetramer, in which all the Zr ions are eight-coordinated, one of the dimer Zr ions is seven-coordinated. This lower coordination does not change in the course of the 10 ps simulation of the dimer in aqueous solution, i.e., a water molecule from the bulk does not fill the vacancy (such a process was observed in, e.g., aluminum chlorohydrate Al_{13} cluster, using the same simulation method and on a similar time scale[23]).

The present work indicates not only that the stacked trimer could exist but also that it may not be a very stable structure, due to a highly strained core consisting of three Zr ions connected by altogether four bridges, where one bridging oxygen atom is shared by all three Zr ions. We observed significant oscillations of one Zr monomer unit with respect to the other two in the first 50 ps of the gas-phase simulation. However, we were not able to detect a similar behavior in solution, due to the confinement of the trimer by the water formed cage around it. We suggest that this instability would be detected in solution as well, if a longer simulation was possible. This argument holds for a seemingly contradictory finding of an X-ray study of the Zr, Ti, and Hf binding to transferrin,[2] a protein that regularly binds iron, which suggests that the trinuclear cluster that is either grown within the protein or bound to the protein cleft is very similar to what we call the stacked trimer. The stability of the cluster within the cleft in this case might be enhanced by the confinement coming from the protein groups; without such confinement, the fluctuations of one of the Zr monomers would be possible.

V. Summary

Despite the importance and widespread use of zirconium hydroxy polynuclear clusters, certain important and basic aspects of their structure and dynamics are not known. This lack of information in some cases limits or even rules out our predictive power with respect to possible functions and applications as well as activities in known applications. By conducting a CPMD simulation study of the two smallest Zr^{4+} polynuclear species, we have provided the necessary basic information on their structure and dynamics in aqueous medium, in which most of the applications are conducted. Our study resulted in a detailed understanding of the structure, including repeating motifs, which will be used for studying larger Zr^{4+} polymers, such as the hexamer, and it reveals for the first time possible configurations of the dimer and trimer. Based on the observed fluctuating internal motion of one Zr monomer with respect to the other two, we postulate that the stacked trimer (structure obtained in our study) has a somewhat unstable structure and might persist only on a short time scale upon its formation, which might be the reason for conflicting experimental reports regarding its existence. Also, our study of the zirconium hexamer[55] reveals that it can be viewed as built from the trimer units. Thus, the trimer could be a transient species in the process of hexamer formation.

Acknowledgment. This research was supported in part by the Petroleum Research Fund administered by the American Chemical Society, in part by the Colgate-Palmolive Company. Computer time from Pittsburgh Supercomputing Center is greatly appreciated. N.R. and V.P. acknowledge support from the H. O. West Foundation and NSF (CHE-0420556) and thank Dr. Z. Liu and Dr. E. Birnbaum for helpful discussions.

References

(1) Ekberg, C.; Kallvenius, G.; Albinsson, Y.; Brown, P. L. *J. Solution Chem.* **2004**, *33*, 47−79.

(2) Alexeev, D.; Zhu, H. Z.; Guo, M. L.; Zhong, W. Q.; Hunter, D. J. B.; Yang, W. P.; Campopiano, D. J.; Sadler, P. J. *Nat. Struct. Biol.* **2003**, *10*, 297−302.

(3) Butler, A. *Nat. Struct. Biol.* **2003**, *10*, 240−241.

(4) Barondeau, D. P.; Getzoff, E. D. *Curr. Opin. Struct. Biol.* **2004**, *14*, 765−774.

(5) Rosenberg, A. H.; Fitzgerald, J. J. *Antiperspirants and Deodorants*; Marcel Dekker, Inc.: New York, Basel, 1999; pp 137−168.

(6) Northrup, C. J. M. J. Immobilization of U.S. defense nuclear wastes using the SYNROC process. In *Scientific Basis for Nuclear Waste Managment*; 1979; Vol. 2, p 265.

(7) Richens, D. T. *The Chemistry of Aqua Ions*; John Wiley & Sons: Chichester, New York, Weinheim, Brisbane, Singapore, Toronto, 1997; pp 207−220.

(8) Baes, C. F.; Mesmer, R. E. *The Hydrolysis of Cations*; John Wiley & Sons: New York, London, Sydney, Toronto, 1976.

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

(10) Pasquarello, A.; Petri, I.; Salmon, P. S.; Parisel, O.; Car, R.; Toth, E.; Powell, D. H.; Fischer, H. E.; Helm, L.; Merbach, A. E. *Science* **2001**, *291*, 856−859.

(11) White, J. A.; Schwegler, E.; Galli, G.; Gygi, F. *J. Chem. Phys.* **2000**, *113*, 4668−4673.

(12) Naor, M. M.; Van Nostrand, K.; Dellago, C. *Chem. Phys. Lett.* **2003**, *369*, 159−164.

(13) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. *Chem. Phys. Chem.* **2005**, *6*, 1745−1749.

(14) Lightstone, F. C.; Schwegler, E.; Hood, R. Q.; Gygi, F.; Galli, G. *Chem. Phys. Lett.* **2001**, *343*, 549−555.

(15) Amira, S.; Spangberg, D.; Zelin, V.; Probst, M.; Hermansson, K. *J. Phys. Chem. B.* **2005**, *109*, 14235−14242.

(16) Ikeda, T.; Hirata, M.; Kimura, T. *J. Chem. Phys.* **2005**, *122*, 024510.

(17) Ramaniah, L. M.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 1587−1591.

(18) Ikeda, T.; Hirata, M.; Kimura, T. *J. Chem. Phys.* **2003**, *119*, 12386−12392.

(19) Lyubartsev, A. P.; Laasonen, K.; Laaksonen, A. *J. Chem. Phys.* **2001**, *114*, 3120−3126.

(20) Marx, D.; Sprik, M.; Parrinello, M. *Chem. Phys. Lett.* **1997**, *273*, 360−366.

(21) Rothlisberger, U.; Klein, M. L. *J. Am. Chem. Soc.* **1995**, *117*, 42−48.

(22) Pophristic, V.; Klein, M. L.; Holerca, M. N. *J. Phys. Chem. A* **2004**, *108*, 113−120.

(23) Pophristic, V.; Balagurusamy, V. S. K.; Klein, M. L. *Phys. Chem. Chem. Phys.* **2004**, *6*, 919−923.

(24) Sillanpaa, A. J.; Paivarinta, J. T.; Hotokka, M. J.; Rosenholm, J. B.; Laasonen, K. E. *J. Phys. Chem. A* **2001**, *105*, 10111−10122.

(25) Clearfield, A.; Vaughan, P. A. *Acta Crystallogr.* **1956**, *9*, 555−558.

(26) Muha, G. M.; Vaughan, P. A. *J. Chem. Phys.* **1960**, *33*, 194−199.

(27) Mak, T. C. W. *Can. J. Chem.* **1968**, *46*, 3491.

(28) Aberg, M. *Acta Chem. Scand. Ser. A - Phys. Inorg. Chem.* **1977**, *31*, 171−181.

(29) Zielen, A. J.; Connick, R. E. *J. Am. Chem. Soc.* **1956**, *78*, 5769.

(30) Johnson, J. S.; Kraus, K. A. *J. Am. Chem. Soc.* **1956**, *78*, 3937−3943.

(31) Angstadt, R. L.; Tyree, S. Y. *J. Inorg. Nucl. Chem.* **1962**, *24*, 913−917.

(32) Veyland, A.; Dupont, L.; Pierrard, J. C.; Rimbault, J.; Aplincourt, M. *Eur. J. Inorg. Chem.* **1998**, 1765−1770.

(33) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703−1710.

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

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

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

(37) The solvation of Zr^{4+} ion will be addressed in a separate publication.

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

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

(40) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

(41) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270−283.

(42) Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03*; Gaussian, Inc.: Wallingford, CT, 2004.

(43) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635−2643.

(44) Nose, S. *J. Chem Phys.* **1984**, *81*, 511.

(45) Nose, S. *Mol. Phys.* **1984**, *52*, 255.

(46) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(47) Rao, N.; Holerca, M. N.; Klein, M. L.; Pophristic, V. *J. Phys. Chem. A* **2007**, *111*, 11395−11399.

(48) Safonov, A. A.; Bagatur’yants, A. A.; Korkin, A. A. *Microelectron. Eng.* **2003**, *69*, 629−632.

(49) Zhao, X.; Ceresoli, D.; Vanderbilt, D. *Phys. Rev. B* **2005**, *71*, 085107.

(50) Foschini, C. R.; Filho, O. T.; Juiz, S. A.; Souza, A. G.; Oliveira, J. B. L.; Longo, E.; Leite, E. R.; Paskocimas, C. A.; Varela, J. A. *J. Mater. Sci.* **2004**, *39*, 1935−1941.

(51) Chen, S. G.; Yin, Y. S.; Wang, D. P. *J. Mol. Struct.* **2004**, *690*, 181−187.

(52) Zhong, W.; Parkinson, J. A.; Parsons, S.; Oswald, I. D. H.; Coxall, R. A.; Sadler, P. *J. Inorg. Chem.* **2004**, *43*, 3561−3572.

(53) Rose, J.; Bruin, T. J. M. D.; Chauveteau, G.; Tabary, R.; Hazemann, J. L.; Proux, O.; Omari, A.; Toulhoat, H.; Bottero, J. Y. *J. Phys. Chem. B* **2003**, *107*, 2910−2920.

(54) Saukkoriipi, j.; Sillanpaa, A.; Laasonen, K. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3785.

(55) Rao, N.; Pophristic, V. To be published.

(56) Zielen, A. J.; Connick, R. E. *J. Am. Chem. Soc.* **1956**, *78*, 5785−5792.

(57) Tribalat, S.; Schriver, L. *Bulletin De La Societe Chimique De France Partie I-Physicochimie Des Systemes Liquides Electrochimie Catalyse Genie Chimique*; 1975; pp 2012−2014.

CT7001094

JCTC Journal of Chemical Theory and Computation

Ab Initio Molecular Dynamics Study of Mg^{2+} and Ca^{2+} Ions in Liquid Methanol

Cristian Faralli,[†] Marco Pagliai,[†] Gianni Cardini*,[†,‡] Vincenzo Schettino[†,‡]

Laboratorio di Spettroscopia Molecolare, Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, 50019 Sesto Fiorentino, Firenze, Italia, and European Laboratory for Nonlinear Spectroscopy (LENS), via Nello Carrara 1, 50019 Sesto Fiorentino, Firenze, Italia

Received August 22, 2007

Abstract: Ab initio Car−Parrinello molecular dynamics simulations have been performed in order to investigate the solvation properties of Mg^{2+} and Ca^{2+} in fully deuterated methanol solution to better understand polarization effects induced by the ions. Charge transfer and dipole moment calculations have been performed to give more detailed insight on the role of the electronic reorganization and its effect on the first solvation shell stability. The perturbation of the methanol H-bond network has been investigated.

Introduction

Simulation studies of the structural and dynamical properties of solutions of ions in polar solvents are of great importance to understand the effects of charged species on the physical and chemical properties of ionic solutions. The presence of ions can strongly perturb the structure of the liquid, and this can have relevant effects on the chemical reactivity in solution. In general a particular role is played by the stability of the first solvation shell although in some cases the perturbation extends farther away from the ion. Despite the large variety of polar solvents of common use in chemistry, only water[1−11] and, to a lower extent, ammonia[4,12,13] have been extensively analyzed from the theoretical and computational point of view. A series of ab initio molecular dynamics and cluster calculations on ions in these solvents[14−20] has been performed showing the importance of polarization interactions in the reproduction of the experimental results.[1,21−24]

At present, methanol, the smallest molecule characterized by both a hydrophobic and a hydrophilic group, is widely used as a solvent, but the number of computational studies on the structural and dynamical properties of the liquid[7,21−29] and its ionic solutions[30−37] is limited. In the past few years the interest on methanol has also grown as a possible fuel cell component.[38−42] Therefore, the comprehension of the liquid methanol properties and its interactions with ions and simple molecules is becoming of paramount importance.

In this paper, we report on ab initio molecular dynamics simulations, within the Car−Parrinello (CPMD)[43−46] formalism, of Mg^{2+} and Ca^{2+} ions in methanol. Particular attention has been paid to the structure of the first solvation shell and to the change of ground-state electronic properties of the ions and of the solvent molecules. This kind of approach has been used with success to study ions in solution[16,30,31,47−59] showing that the most relevant effects are concerned with the first solvation shell. The nature of the interactions that stabilize the first solvation shell of the Mg^{2+} and Ca^{2+} ions in methanol has been interpreted in terms of charge transfer and polarization, confirming the stabilization model proposed in the case of the Li^+,[30] Na^+, and K^+ ions[31] in the same solvent.

Computational Details

The simulations have been performed with the CPMD code[43,46] in cubic boxes of 12.05 Å and 13.99 Å side with periodic boundary conditions using 25 and 40 methanol molecules, respectively, and one ion. The initial configuration with 25 molecules sample has been constructed starting from the last configuration extracted from a previous Car−Parrinello simulation of Li^+ in methanol[30] with the simple ion substitution. The initial configurations for the larger systems have been taken from the last configuration obtained performing a preliminary classical simulation (>100 ps)

* Corresponding author e-mail: gianni.cardini@unifi.it.

[†] Università di Firenze,.

[‡] European Laboratory for Nonlinear Spectroscopy (LENS).

10.1021/ct700209v CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/18/2007

using semiempirical potentials with Lennard-Jones parameters taken from the literature.[60]

After a thermalization at 300 K by velocity scaling (∼1 ps), the equations of motions have been integrated with a time step of 5 au (∼0.12 fs) in the NVE ensemble storing atomic coordinates and velocities for the subsequent analysis. The simulation time has been of ∼16 ps and ∼9 ps for Mg^{2+} and Ca^{2+}, respectively, in the 25 solvent molecules samples and of ∼11 ps for the samples with 40 methanol molecules. The computational protocol adopted in previous works[30,31,58] has shown that the time scale is sufficient to accurately reproduce the structural properties. The correct conservation of the energy, a sensitive parameter in CPMD charged systems, has been monitored during the whole simulation run.

Most of the analysis reported here have been performed on trajectories from simulations of the larger samples.

The deuterium has been used instead of hydrogen to allow for a larger time step. Density functional calculations in the generalized gradient approximation (GGA) have been performed using the BLYP[61,62] exchange correlation functional. A ficticious electronic mass of 800 au has been adopted to keep the system on the Born−Oppenheimer surface. The plane wave (PW) expansion has been truncated at 70 Ry.

Martins-Troullier[63] pseudopotentials (MT) have been used along with the Kleinman-Bylander[64] decomposition for the C, H, and O atomic species. For the calcium ion preliminary simulations with 25 methanol molecules have been performed adopting either a MT semicore pseudopotential (considering as core the 1s, 2s, and 2p electrons) or a Goedecker semicore pseudopotential (SG)[65,66] in order to analyze possible effects of the pseudopotential choice. The results of the two simulations showed very similar structural properties as reported in the Supporting Information (Figure 1-S). In the larger samples, for both calcium and magnesium ions, Goedecker[65,66] type pseudopotentials have been adopted as in previous works on monovalent cations in methanol.[30,31]

The reliability of the pseudopotentials and of the computational approach chosen for the simulations and the subsequent analysis has been confirmed by the convergence of the structural data with plane waves cutoff as reported in Table I-S of the Supporting Information. A good binding energy[67] convergence at 70 Ry has been obtained. Considering the higher extension of the basis set used in this work with respect to those present in literature,[68] adopting the B3LYP/6-31+G* level of theory calculations, the agreement with the literature values is fairly good. The "all electrons" calculations show a lower binding energy value for Mg^{2+} with respect to PWs expansion, with an opposite trend for Ca^{2+}, but in both cases the differences are lower than 2 kcal mol^{-1}.

Atoms in molecules (AIM)[69−71] and maximally localized Wannier function centers (WFCs) analysis[72,73] have been performed averaging over equi-spaced configurations in the samples with 40 methanol molecules, every 0.08 ps for Mg^{2+} and 0.09 ps for Ca^{2+}.

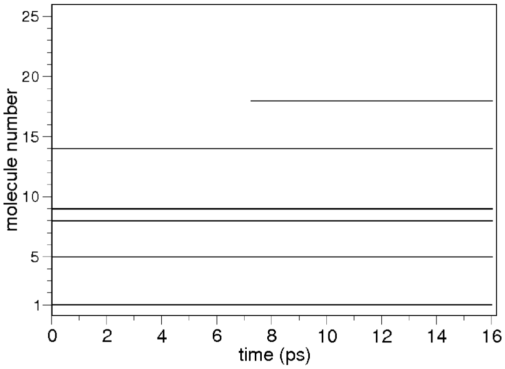


Figure 1. A dot is reported at each time step for a molecule in the first solvation shell.

Results and Discussion

The systems have been initially simulated in samples with 25 methanol molecules. In the case of the Mg^{2+} a peculiar behavior has been noticed. A 5-fold coordination has been observed during the initial 7.3 ps of the run (see Figure 1) with a square pyramidal basis coordination geometry (see Figure 2a). Subsequently the number of methanol molecules around the ion rises up abruptly to a stable 6-fold octahedral coordination (Figures 1 and 2b).

In order to explain this behavior, the energy of the optimized geometry for the two configurations (extracted before and after the coordination number change) has been computed for isolated clusters with "all electrons" calculations, using the BLYP functional and the 3-21+G** basis set. The clusters coordinates are reported in Tables II-S and III-S of the Supporting Information, and the first shell configurations are shown in Figure 2a,b. The results show a higher stability of the 6-fold coordinated cluster with a difference in the binding energy of 29.95 kJ mol^{-1}. This value is about 1 order of magnitude higher than the thermal energy at 300 K (2.49 kJ mol^{-1}) and explains the observed stability of the 6-fold coordinated ion once it is formed. The initial 5-fold configuration can be attributed to the selected starting configuration and to a likely too short thermalization run (∼1 ps) with respect to the cage relaxation time.

It is interesting to note that the salient structural data for the two parts of the simulation are not particularly different, as in Figure 2-S of the Supporting Information. In particular, comparison of the two partial pair distribution functions with the total shows that the first peak position is not appreciably affected by the variation of the coordination number.

In Figure 3 the pair radial distribution function for the Mg−O and Ca−O distances, together with their integration number, are reported for samples with 40 methanol molecules and compared with the system containing 25 solvent molecules.

It is evident that, for both ions, the sample dimension only affects the second solvation shell that is slightly better defined in the larger sample, showing clearly that it is formed by 6 molecules. In the case of Ca^{2+} a small effect on the height and width of the first peak can also be noted implying a greater rigidity of the first solvation shell in the smaller simulation box. The stability of the second shell is higher for the larger samples as it can be inferred from the slightly deeper second minimum.

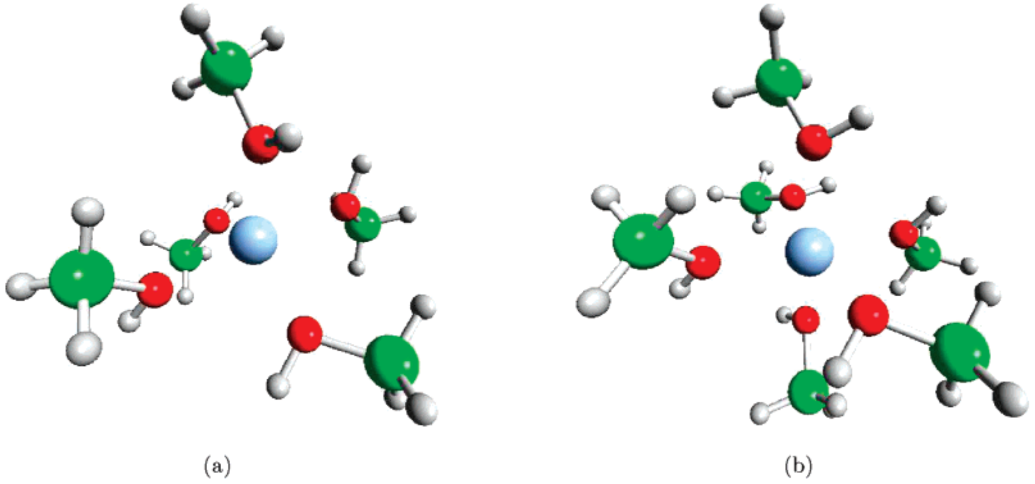


Figure 2. Mg ion and its nearest neighbors: (a) configuration extracted from the first part of simulation (pentacoordination) and (b) configuration referred to the second part of simulation.

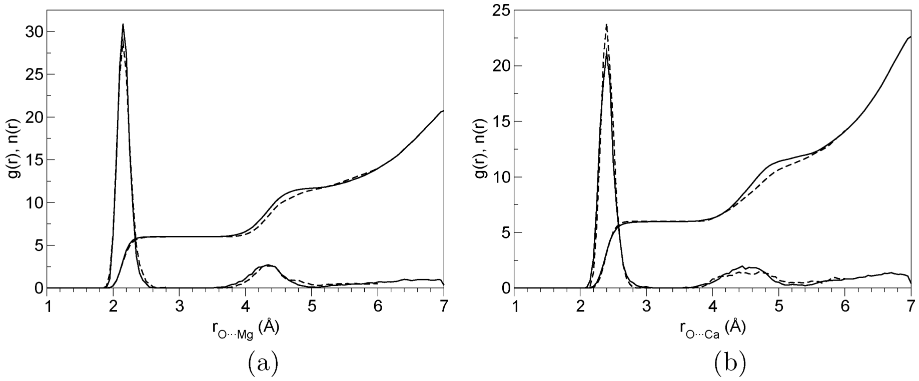


Figure 3. Pair radial distribution functions and integration numbers of Mg^{2+} (a) and Ca^{2+} (b) with 25 methanol molecules (dashed lines) and with 40 methanol molecules (full lines), respectively.

Table 1. Salient Structural Data (Distances in Å) for Mg^{2+} and Ca^{2+} Solutions with 25 and 40 Solvent Molecules[a]

	O···M^{2+}	cutoff	*n*(*r*)
Mg^{2+} (25)	2.15	3.00	5.6 (5 or 6)
Mg^{2+} (40)	2.15	3.00	6.0
Radnai et al.[74]	2.068		5.95
Tamura et al.[75]	2.00	2.5−3	6.0
Ca^{2+} (25)	2.40	3.25	6.0
Ca^{2+} (40)	2.40	3.45	6.0
Megyes et al.[76,77]	2.39		6.0

[a] The coordination number, *n*(*r*), has been computed at the cutoff distance. The data are compared with experimental results.

Table 1 reports the position of the first peak in the radial distribution function and the integration number.

It can be seen that the cutoff distance has no effect on the coordination number due to the fact that the first minimum in the *g*(*r*) is widespread. X-ray diffraction studies[74] locate the first peak position at 2.068 Å with a “relatively rigid octahedral” cage thus proposing a 6-fold coordination. Subsequent studies, supported by molecular dynamics simulations,[75] confirmed these findings although with a first peak position at shorter distance (2.00 Å) than the X-ray result. In the present calculation the first peak position for Mg^{2+} solutions is found at a slightly larger distance (2.15 Å). The results of the present simulation are in full agreement with experiments[76,77] in the case of the Ca^{2+} ion.

For both ions the residence time of the methanol molecules in the first solvation shell is longer than the simulation time, and no exchange of methanol molecules has been observed between the first and second solvation shell, as can also be argued by the flat and deep minimum in the pair radial distribution functions. A similar behavior has been reported for water solution where many of these dications are surrounded by a rigid first solvation shell that shows a slow exchange of water molecules with the second shell.[14,15,78−80] Earlier, diffusion coefficient calculations and solvation simulations reported a very long lifetime for water molecules in the first solvation shell around Mg^{2+}, falling in the range of hundreds of picoseconds.[81,82]

The small amplitude of motion in the cage is well evident from the pair distribution functions (Figure 3) characterized by a very sharp first peak. This is further emphasized by the angular distribution function reported in Figure 3a-S along with the spatial distribution function[83−86] of Figure 3b-S obtained from the configurational space spanned by the ion considering a methanol molecule of the first solvation shell as the reference system. As expected, the oxygen lone pairs of the methanol molecules are steadily oriented in the

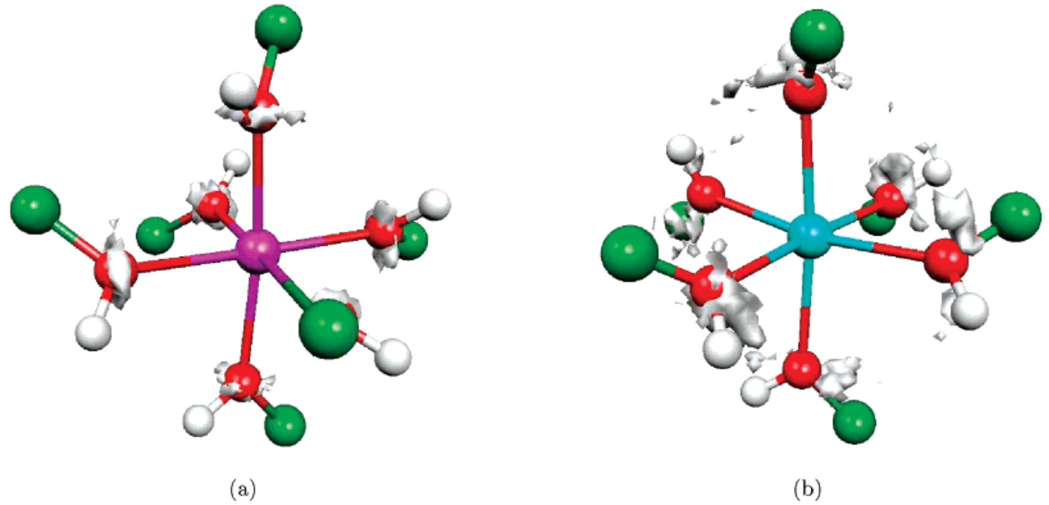


Figure 4. Spatial distribution functions for the first solvation shell of Mg^{2+} (a) and Ca^{2+} (b) in the system with 40 methanol molecules. The isosurface represents the 13% and the 16% of the maximum value for Mg^{2+} and Ca^{2+}, respectively. The methyl groups have been represented by the green spheres.

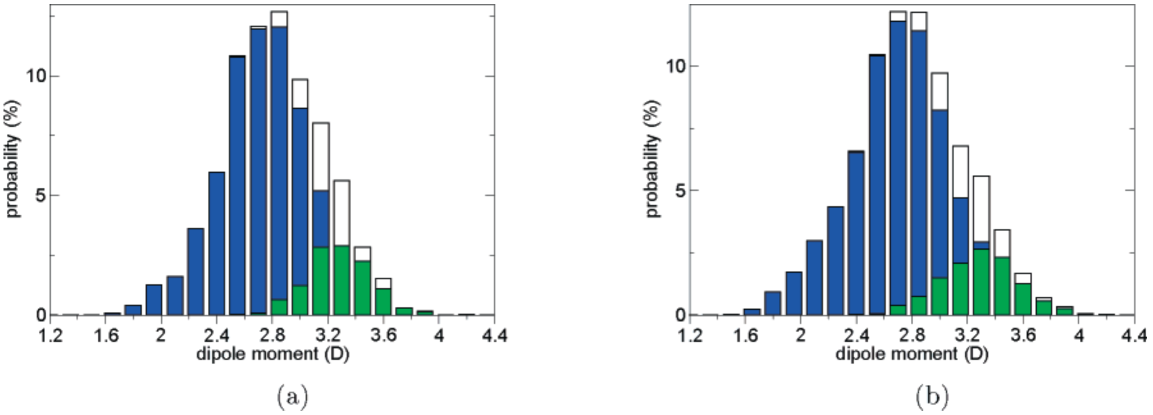


Figure 5. Dipole moment for methanol molecules with Mg^{2+} (a) and Ca^{2+} (b). The green bars refer to the dipole moment of the first shell molecules, whereas the blue bars describe the external molecules contribution. The average dipole moment of the whole solution is represented by the white bars.

Table 2. Average Dipole Moment Values (in Debye, D) and Relative Standard Deviation for the Solution ($<\mu>_{tot}$), for the First Shell Molecules Contribution ($<\mu>_{fs}$), and for the External Molecules ($<\mu>_{ext}$)

	$<\mu>_{tot}$	$<\mu>_{fs}$	$<\mu>_{ext}$
Mg^{2+} 25	2.9 ± 0.4	3.4 ± 0.4	2.7 ± 0.3
Mg^{2+} 40	2.8 ± 0.4	3.3 ± 0.2	2.7 ± 0.3
Ca^{2+} 25	2.8 ± 0.4	3.2 ± 0.3	2.7 ± 0.3
Ca^{2+} 40	2.8 ± 0.4	3.3 ± 0.3	2.7 ± 0.4

direction of Ca^{2+} with the COCa and HOCa tilt angles around 120° and 127°, respectively, with a small amplitude of motion around the ion.

To better characterize the structure of the cage in the larger sample salient average distances and angles are reported in Table IV-S. A pictorial view of the first solvation shell is shown in Figure 4 where the motion amplitude of the oxygen atoms around the ion is displayed. The spanned configurational space is found to be strictly localized around the vertices of an octahedron, particularly for Mg^{2+}, as can also be argued from the lower dispersion of the data.

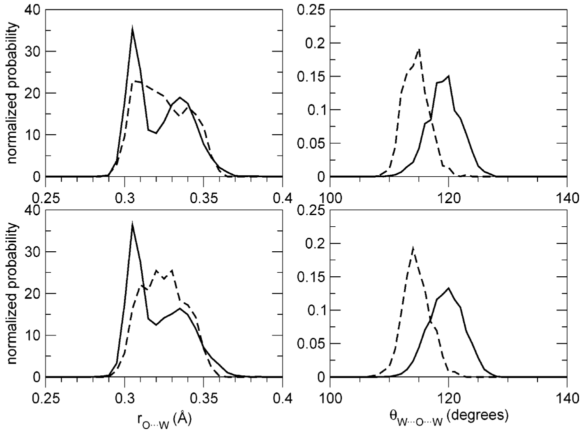


Figure 6. Radial distribution function of the oxygen-WFCs average distance ($r_{O\cdots W}$) and angular distribution function of the angle between the WFCs ($\theta_{W\cdots O\cdots W}$) for Mg^{2+} with 40 methanol molecules (top) and for Ca^{2+} with 40 methanol molecules (bottom). The dashed lines refer to the first shell molecules contribution.

The perturbation on the solvent structure, due to the presence of the ion, has been evaluated in terms of electronic

Table 3. Hydrogen Bond Network Characterization[a]

	Mg^{2+}		Ca^{2+}			Mg^{2+}		Ca^{2+}	
	fs	tot	fs	tot		fs	tot	fs	tot
f_0	100	18	100	22	g_0	100	1	99	3
f_1	0	67	0	62	g_1	0	18	1	23
f_2	0	15	0	16	g_2	0	65	0	58
f_3					g_3	0	16	0	16
$<n_f>$	0.00	0.98	0.00	0.94	$<n_g>$	0.00	1.95	0.01	1.85

[a] f_i represents the percentage of methanol molecules whose oxygen atom is involved in *i* H-bonds, and $<n_f>$ is the average number of received H-bonds per molecule; g_i is the percentage of methanol molecules that totally form *i* H-bonds, and $<n_g>$ is the average number of total H-bonds per molecule. The first shell (fs) molecules contribution has been put into evidence.

properties that illustrate the differences from the pure solvent.[25] The polarization effects are evidenced by the dipole moment computed through the maximally localized WFCs and shown in Figure 5.

The ion perturbation mainly affects the neighboring molecules that are highly polarized as it is seen from the change of the average total dipole moment ($\Delta\mu \sim 0.4$ D). In turn the dipole moment of the outer molecules approaches the value of the pure liquid (2.6 D)[25] remarking the weaker perturbation at long range. These results are summarized in Table 2 where it can be noted that the contribution to the dipole moment does not depend on the system size. In a recent paper,[31] the polarization provided by monovalent cations on the surrounding methanol molecules was found weaker. The stronger polarization due to these divalent cations with respect to monovalent ones has been also reported in water solution.[59]

For alkaline ions the size of the ions increases going from Li^+ to K^+, and, consequently, as expected, the induced dipole moment of the solvent molecules decreases. This is particularly evident for the first solvation shell molecules.[31] For magnesium and calcium ions a different trend can be observed: the longer O−Ca distance does not yield a weaker polarization effect with respect to magnesium ion, and the perturbation on the dipole moment values is similar for both ions.

Radial $r_{O\cdots W}$ and angular $\theta_{W\cdots O\cdots W}$ distributions of the WFCs of the oxygen lone pairs have been investigated to better understand the increase of the dipole moment of the first solvation shell molecules.[87] These are reported in Figure 6 showing separately the contribution of the first solvation shell molecules.

A different shape in the distribution of the $r_{O\cdots W}$ distances and a lower value in the $\theta_{W\cdots O\cdots W}$ angles are observed for the first shell molecules, whereas no change occurs between the oxygen and the WFCs attributed to the O−H and O−C covalent bonds (not reported). For the $\theta_{W\cdots O\cdots W}$ angle a smaller value can be observed for the lone-pair WFCs of the first shell molecules both in methanol and water solution,[16,50,87,88] while a different behavior is present in the $r_{O\cdots W}$ distance in the two solvents. A double peak in the distribution of $r_{O\cdots W}$ is found for the external molecules in methanol, while a symmetrical distribution has been found for the molecules directly solvating the ion. In the bulk, where no coordinative constrain is imposed, some methanol molecules are directionally H-bonded[25] through a single WFC. The methanol H-bonded lone pairs are less contracted on the oxygen than the noninteracting lone pairs providing a splitting of the peak. No H-bond network is permitted between the molecules of the first solvation shell. This can be attributed to the steric hindrance of the CH_3 group. The first shell methanol molecules interact with the ion through both the WFCs that are therefore not anymore available to accept hydrogen atoms from other methanol molecules. This is summarized in Table 3 (left panel) where the percentage of solvent molecules that accept H-bond formation, via oxygen atom, is reported along with the average number of accepted H-bonds per molecule. It can be noticed that no methanol molecule in the first solvation shell is a H-bond acceptor.

The table also reports (right panel) the distribution of the total hydrogen bonds in the system, showing that the majority of methanol molecules are involved in two hydrogen bonds, a result in agreement with finding in the pure solvent.[25,89,90]

Further insight on the solvent reorganization produced by the ion is obtained considering the angle θ_μ between the dipole moment vector ($\vec{\mu}$) and the oxygen-ion interaction axis (see Figure 7).[75,91]

It can be seen from the figure that for the first shell molecules θ_μ is quite tightly peaked around 18°, indicating a rigid structure of the solvation shell. For the outer molecules the distribution is very shallow. A similar behavior has been generally observed in water solutions[47,49,59,92−95] and only rarely in other solvents.[32] This behavior can be further enlightened considering the variation of the dipole moment orientation and its standard deviation as a function of the distance from the central ion (Figure 8).[75,91−93]

Neglecting the 3−4 Å range, where the statistics are rather poor, it can be seen that, up to 5 Å, θ_μ increases smoothly, and the deviations from the average value are small implying that there is a preferential orientation of the dipoles in the first and even in the second shell. Above 5 Å a higher disorder in the bulk of the solution is evident. Similar results have been obtained for Mg^{2+} (see Figure 5-S).

The charge-transfer analysis on the ions has been performed using the AIM approach proposed by Bader.[70] This method also allows the evaluation of the amount of the charge transfer as a function of the distance between the ion and the surrounding solvent molecules as it is depicted for Mg^{2+} with 40 methanol molecules in Figure 9.

The electronic charge transfer on the ions is 0.221 ± 0.003 e^- for Mg^{2+} and 0.347 ± 0.008 e^- for Ca^{2+}. The same trend was also observed in water solutions.[59] A smaller electronic

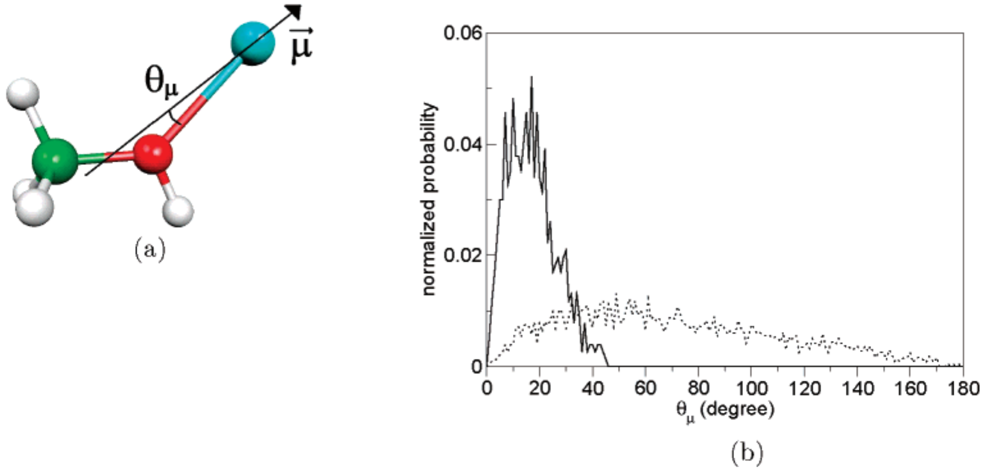


Figure 7. (a) Definition of the θ_μ angle between the dipole vector ($\vec{\mu}$) direction of the methanol molecules and the Ca−O axis. (b) Distribution function of θ_μ for the system with 40 molecules. The full line refers to the first shell contribution. The dotted line represents all other external molecule.

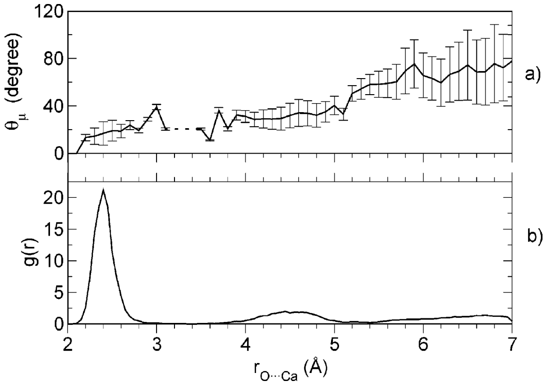


Figure 8. (a) Standard deviation on the average value of the θ_μ angle as a function of the distance from the calcium ion in the system with 40 methanol molecules. (b) O−Ca pair distribution function (same as Figure 3b with full lines).

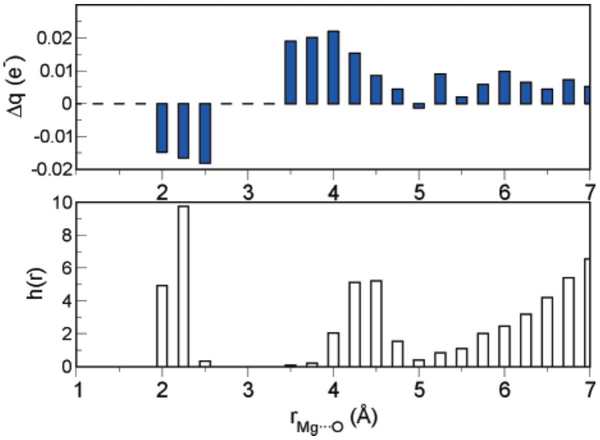


Figure 9. Charge-transfer distribution $\Delta q(e^-)$ as a function of the distance to the ion (upper panel) and not normalized O−Mg distribution function (lower panel) for Mg^{2+} with 40 methanol molecules.

displacement was observed for alkali metal cations together with a weaker polarization effect on the first shell molecules.[30,31] The higher value for the calcium ion is due to its greater softness with respect to magnesium as expected in going down along the group in the periodic table.[96] The same number of valence shell electrons and the same charge on Mg^{2+} and Ca^{2+} are distributed in a different atomic volume. The higher ionic radius[97,98] of Ca^{2+} implies a difference in hardness,[99,100] namely the resistance of the chemical potential to change the number of electrons.[96] Ca^{2+} can receive a greater charge amount from the first shell molecules that become very positively charged. The charge transfer from the second shell to the first is not sufficient to balance the charge transferred from the methanol molecules of the first solvation shell to the ion.

Conclusions

Ab initio CPMD calculations have been performed on solutions of methanol with Mg^{2+} and Ca^{2+} in order to investigate the reorganization effects on a protic solvent due to the presence of charged species. The reliability of the method has been stated by comparison with “all electrons” calculations with a localized Gaussian basis set, showing good agreement in the ion−methanol interaction. The first solvation shell properties have been analyzed and compared to the bulk solvent molecules. The structure of the solvent has been investigated using distribution functions and the electronic properties through the AIM population analysis and the maximally localized WFCs. The box size effects have been explored, and no evident consequence has been found on the first solvation shell.

Similar structural and electronic reorganization is induced on the solvent by the two ions. A stable octahedral coordination and a high polarization effect on the molecules of the first solvation shell have been observed. Analysis of the dipole moment vector has shown a preferential orientation up to 5 Å far from the Ca^{2+} with an influence on the organization of the second shell molecules as well. The characterization of the hydrogen bond network has shown a different trend with respect to that observed in water solution[16,50,87,88] without any solvent molecule in the first solvation shell behaving as a H-bond acceptor. Electronic charge-transfer analysis has confirmed the stabilization of the first solvation shell due to electrostatic interactions as discussed in the literature.[30,31]

Acknowledgment. This work has been supported by the Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica (MIUR).

Supporting Information Available: Pair radial distribution function and an octahedral rearrangement of the cage around the ion (Figure 1-S), structural parameters for methanol−Mg^{2+} and methanol−Ca^{2+} clusters (Table I-S), Cartesian coordinates (Tables II-S and III-S), MgO pair radial distribution function and its running integration numbers for all the runs (Figure 2-S), angular and spatial distribution functions (Figure 3-S), structural parameters (Table IV-S and Figure 4-S), and standard deviation on the average value of the θ_μ (Figure 5-S).This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Chem. Phys. B* **2004**, *108*, 12990−12998.

(2) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.

(3) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. *J. Phys. Chem. B* **2006**, *110*, 3685−3691.

(4) Liu, Y.; Tuckerman, M. E. *J. Phys. Chem. B* **2001**, *105*, 6598−6610.

(5) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308−3311.

(6) Silvestrelli, P.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 3572−3580.

(7) Ladanyi, B. M.; Skaf, M. S. *Ann. Rev. Phys. Chem.* **1993**, *44*, 335−368.

(8) Gubskaya, A. V.; Kusalik, P. *J. Chem. Phys.* **2002**, *117*, 5290−5302.

(9) Luzar, A.; Chandler, D. *Phys. Rev. Lett.* **1996**, *76*, 928−931.

(10) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219−260.

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

(12) Boese, A. D.; Chandra, A.; Martin, J. M. L.; Marx, D. *J. Chem. Phys.* **2003**, *119*, 5965−5980.

(13) Diraison, M.; Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *111*, 1096−1103.

(14) Obst, S.; Bradaczek, H. *J. Phys. Chem.* **1996**, *100*, 15677−15687.

(15) Schwenk, C. F.; Loeffler, H.; Rode, B. *Chem. Phys. Lett.* **2001**, *349*, 99−103.

(16) Bakó, I.; Hutter, J.; Pálinkás, G. *J. Chem. Phys.* **2002**, *117*, 9838−9843.

(17) Merrill, G. N.; Webb, S.; Bivin, D. *J. Phys. Chem. A* **2003**, *107*, 386−396.

(18) Sprik, M.; Impey, R. W.; Klein, M. L. *Phys. Rev. Lett.* **1986**, *56*, 2326−2329.

(19) Martyna, G. J.; Klein, M. L. *J. Chem. Phys.* **1992**, *96*, 7662−7671.

(20) Mundy, C. J.; Kuo, I.-F. W. *Chem. Rev.* **2006**, *106*, 1282−1304.

(21) Tsuchida, E.; Kanada, Y.; Tsukada, M. *Chem. Phys. Lett.* **1999**, *311*, 236−240.

(22) Morrone, J. A.; Tuckerman, M. E. *Chem. Phys. Lett.* **2003**, *370*, 406−411.

(23) Morrone, J. A.; Tuckerman, M. E. *J. Chem. Phys.* **2002**, *117*, 4403−4413.

(24) Handgraaf, J.-W.; Meijer, E. J. *J. Chem. Phys.* **2004**, *121*, 10111−10119.

(25) Pagliai, M.; Cardini, G.; Righini, R.; Schettino, V. *J. Chem. Phys.* **2003**, *119*, 6655−6662.

(26) Handgraaf, J.; van Erp, T.; Meijer, E. *Chem. Phys. Lett.* **2003**, *367*, 617−624.

(27) Haughney, M.; Ferrario, M.; McDonald, I. R. *Mol. Phys.* **1986**, *88*, 849−853.

(28) Haughney, M.; Ferrario, M.; McDonald, I. R. *J. Phys. Chem.* **1987**, *91*, 4934−4940.

(29) Jorgensen, W. L. *J. Am. Chem. Soc.* **1980**, *102*, 543−549.

(30) Pagliai, M.; Cardini, G.; Schettino, V. *J. Phys. Chem. B* **2005**, *109*, 7475−7481.

(31) Faralli, C.; Pagliai, M.; Cardini, G.; Schettino, V. *Theor. Chem. Acc.* **2007**, *118*, 417−423.

(32) Impey, R. W.; Sprik, M.; Klein, M. L. *J. Am. Chem. Soc.* **1987**, *109*, 5900−5904.

(33) Jorgensen, W. L.; Bigot, B.; Chandrasekhar, J. *J. Am. Chem. Soc.* **1982**, *104*, 4584−4591.

(34) Chandrasekhar, J.; Jorgensen, W. L. *J. Chem. Phys.* **1982**, *77*, 5080−5089.

(35) Sesé, G.; Padró, J. A. *J. Chem. Phys.* **1998**, *108*, 6347−6352.

(36) Islam, M. S.; Pethrick, R. A.; Pugh, D. *J. Phys. Chem. A* **1998**, *102*, 2201−2208.

(37) Masella, M.; Cuniasse, P. *J. Chem. Phys.* **2003**, *113*, 1866−1873.

(38) Narayanan, S.; Gottesfelf, S.; Zawodzinski, T. *Direct Methanol Fuel Cells*; Electrochemical Society: November 2001.

(39) Whitacre, J.; Valdez, T.; Narayanan, S. *J. Electrochem. Soc.* **2005**, *152*, A1780−A1789.

(40) Seo, M.; Yun, Y.; Lee, J.; Tak, Y. *J. Power Sources* **2006**, *159*, 59−62.

(41) Yang, Y.; Liang, Y. C. *J. Power Sources* **2007**, *165*, 185−195.

(42) Wang, Z.; Yin, G.; Shao, Y.; Yang, B.; Shi, P.; Feng, P. *J. Power Sources* **2007**, *165*, 9−15.

(43) *CPMD V 3.9 Copyright IBM Corp 1990−2004*; Copyright MPI für Festkörperforschung: Stuttgart, Germany, 1997−2001.

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

(45) Tse, J. S. *Ann. Rev. Phys. Chem.* **2002**, *53*, 249−290.

(46) Hutter, J.; Iannuzzi, M. *Z. Kristallogr.* **2005**, *220*, 549−551.

(47) White, J.; Schwegler, E.; Galli, G.; Gygi, F. *J. Chem. Phys.* **2000**, *113*, 4668−4673.

(48) Ikeda, T.; Hirata, M.; Kimura, T. *J. Chem. Phys.* **2003**, *119*, 12386−12392.

(49) Ramaniah, L. M.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 1587−1591.

(50) Lightstone, F. C.; Schwegler, E.; Hood, R. Q.; Gygi, F.; Galli, G. *Chem. Phys. Lett.* **2001**, *343*, 549−555.

(51) Marx, D.; Sprik, M.; Parrinello, M. *Chem. Phys. Lett.* **1997**, *273*, 360−366.

(52) Naor, M. M.; Nostrand, K. V.; Dellago, C. *Chem. Phys. Lett.* **2003**, *369*, 159−164.

(53) Blumbergerer, J.; Sprik, M. *J. Phys. Chem. B* **2004**, *108*, 6529−6535.

(54) Blumberger, J.; Bernasconi, L.; Tavernelli, I.; Vuilleumier, R.; Sprik, M. *J. Am. Chem. Soc.* **2004**, *126*, 3928−3938.

(55) Lyubartsev, A. P.; Laasonen, K.; Laaksonen, A. *J. Chem. Phys.* **2001**, *114*, 3120−3126.

(56) Vuilleumier, R.; Sprik, M. *J. Chem. Phys.* **115**, *8*, 3454−3468.

(57) Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **1995**, *103*, 150−161.

(58) Faralli, C.; Pagliai, M.; Cardini, G.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14923−14928.

(59) Krekeler, C.; Delle Site, L. *J. Phys.: Condens. Matter* **2007**, *19*, 192101.

(60) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276−1284.

(61) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

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

(63) Troullier, N.; Martins, J. *Phys. Rev. B* **1991**, *43*, 1993−2006.

(64) Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425−1428.

(65) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703−1710.

(66) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641−3662.

(67) Mercero, J.; Mujika, J.; Matxain, J.; Lopez, X.; Ulgade, J. *Chem. Phys.* **2003**, *295*, 175−184.

(68) Dudev, T.; Lim, C. *J. Phys. Chem. A* **1999**, *103*, 8093−8100.

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

(70) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893−928.

(71) Szefczyk, B.; Sokalski, W. A.; Leszczynski, J. *J. Chem. Phys.* **2002**, *117*, 6952−6958.

(72) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847−12862.

(73) Silvestrelli, P. L.; Marzari, N.; Vanderbilt, D.; Parrinello, M. *Solid State Commun.* **1998**, *107*, 7−11.

(74) Radnai, T.; Kálmán, E.; Pollmer, K. *Z. Naturforsch., A: Phys. Sci*. **1984**, *39A*, 464−470.

(75) Tamura, Y.; Spohr, E.; Heinzinger, K.; Pálinkás, G.; Bakó, I. *Ber. Bunsenges Phys. Chem.* **1992**, *96*, 147−158.

(76) Megyes, T.; Grósz, T.; Radnai, T.; Bakó, I.; Pálinkás, G. *J. Phys. Chem. A* **2004**, *108*, 7261−7271.

(77) Megyes, T.; Bálint, S.; I. Bakó, T. G.; Radnai, T.; Pálinkás, G. *Chem. Phys.* **2006**, *327*, 415−426.

(78) Raugei, S.; Klein, M. L. *J. Chem. Phys.* **2002**, *116*, 196−202.

(79) Ohtaki, H.; Radnai, T. *Chem. Rev.* **1993**, *93*, 1157−1204.

(80) Richens, D. T. *The Chemistry of Aqua Ions*; Wiley: 1997.

(81) Masia, M.; Rey, R. *J. Chem. Phys.* **2005**, *122*, 094502.

(82) Jiao, D.; King, C.; Grossfield, A.; Darden, T.; Ren, P. *J. Phys. Chem. B* **2006**, *110*, 18553−18559.

(83) Svishchev, I. M.; Kusalik, P. G. *J. Chem. Phys.* **1993**, *99*, 3049−3058.

(84) Svishchev, I. M.; Kusalik, P. G. *J. Chem. Phys.* **1994**, *100*, 5165−5171.

(85) Khalack, J. M.; Lyubartsev, A. P. *J. Chem. Phys.* **2005**, *109*, 378−386.

(86) De la Peña, L. H.; Kusalik, P. *J. Am. Chem. Soc.* **2005**, *127*, 5246−5251.

(87) Ikeda, T.; Boero, M.; Terakura, K. *J. Chem. Phys.* **2007**, *126*, 034501.

(88) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. *Chem. Phys. Chem.* **2005**, *6*, 1745−1749.

(89) Sun, W.; Chen, Z.; Huang, S.-Y. *Fluid Phase Equilib.* **2005**, *238*, 20−25.

(90) Suresh, S. J.; Prabhu, A.; Arora, A. *J. Chem. Phys.* **2007**, *126*, 134502.

(91) Marx, D.; Heinzinger, K.; Pálinkás, G.; Bakó, I. *Z. Naturforsch., A: Phys. Sci*. **1991**, *46A*, 887−897.

(92) Palinkas, G.; Heinzinger, K. *ACH - Models Chem.* **1995**, *132*, 5−29.

(93) Tongraar, A.; Liedl, K.; Rode, B. *J. Phys. Chem. A* **1998**, *102*, 10340−10347.

(94) Brodskaya, E.; Lyubartsev, A. P.; Laaksonen, A. *J. Chem. Phys.* **2006**, *116*, 7879−7892.

(95) Tongraar, A.; Rode, B. *Chem. Phys. Lett.* **2001**, *346*, 485−491.

(96) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512−7516.

(97) Shannon, R. D.; Prewitt, C. *Acta Crystallogr., Sect. B: Struct. Sci.* **1969**, *B25*, 925−946.

(98) Shannon, R. D. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1976**, *A32*, 751−767.

(99) Fuentealba, P.; Parr, R. G. *J. Chem. Phys.* **1991**, *94*, 5559−5564.

(100) Ghanty, T.; Ghosh, S. *J. Phys. Chem.* **1994**, *98*, 9197−9201.

CT700209V

Journal of Chemical Theory and Computation

'Dynamic Distance' Reaction Coordinate for Competing Bonds: Applications in Classical and Ab Initio Simulations

Christian Burisch,[†,‡] Phineus R. L. Markwick,[†,§] Nikos L. Doltsinis,[||,#] and Jürgen Schlitter*,[‡]

Lehrstuhl für Biophysik, Ruhr-Universität Bochum, ND 04, 44780 Bochum, Germany, Unité de Bioinformatique Structurale, Institut Pasteur, CNRS URA 2185, 25-28 Rue du Dr. Roux, 75015 Paris, France, and Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany

Received July 11, 2007

Abstract: A versatile reaction coordinate, the "dynamic distance", is introduced for the study of reactions involving the rupture and formation of a series of chemical bonds or contacts. The dynamic distance is a mass-weighted mean of selected distances. When implemented as a generalized constraint, the dynamic distance is particularly suited for driving activated processes by controlled increase during a simulation. As a single constraint acting upon multiple degrees of freedom, the sequence of events along the resulting reaction pathway is determined unambiguously by the underlying energy landscape. Free energy profiles can be readily obtained from the mean constraint force. In this paper both theoretical aspects and numerical implementation are discussed, and the unique and diverse properties of this reaction coordinate are demonstrated using three examples: In the framework of Car−Parrinello molecular dynamics, we present results for the prototypical double proton-transfer reaction in formic acid dimer and the photocycle of the guanine−cytosine DNA base pair. As a classical mechanical example, the opening of the binding pocket of the enzyme rubisco is analyzed.

1. Introduction

A reaction coordinate (RC) provides a measure of the progress of an activated process, such as a chemical reaction, from an initial reactant to a final product state. The RC is usually defined in advance without prior knowledge of the actual pathway (or pathways), and so the choice of the coordinate is guided by a preliminary postulated picture of the reaction. Nevertheless, the reaction coordinate represents a valuable tool to enforce a transition away from the reactant state or toward the product state. This 'coordinate driving' approach is one of the valid methods for pathway search reviewed recently[1] which also provides a parametrization of the associated complex free energy surface. To be successful, a suitable coordinate must be constructed specifically for the problem at hand. Numerous examples can be found in the literature ranging from simple distance coordinates[2] and weighted combinations of distances[3,4] to more abstract coordination numbers[5] or even energy.[6,7] Activated processes or reactions can be driven by implementing auxiliary restraint potentials in the framework of umbrella sampling[8] or by applying holonomic constraints, as demonstrated in recent applications of targeted molecular dynamics.[9] Both these approaches enable the computation of free energy profiles using recently refined techniques[10,11] and can be applied to the study of a diverse range of systems from elementary chemical reactions to large scale conformational transitions in biological macromolecules.[12] Nevertheless, any reaction coordinate can (and usually does) provide a simplistic picture of a reaction, being a compromise between free exploration

* Corresponding author e-mail: juergen.schlitter@rub.de.

[†] These authors contributed equally to this work.

[‡] Lehrstuhl für Biophysik, Ruhr-Universität Bochum.

[§] Institut Pasteur.

[||] Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum.

[#] Present address: Department of Physics, King's College London, London WC2R 2LS, United Kingdom.

10.1021/ct700170t CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/08/2007

of the pathway through phase space and optimal accuracy of free energy.[13]

In this paper we present a novel reaction coordinate, the ‘dynamic distance’, which has been specifically designed for the study of reactions involving the rupture and/or formation of chemical bonds or contacts, such as salt bridges. This flexible reaction coordinate, formulated as a mass-weighted mean of selected interatomic distances, drives the activated process without influencing the sequence or mechanism of the events, such that the resulting reaction pathway is determined only by the underlying potential energy landscape. As a single coordinate constructed from multiple internal degrees of freedom, the dynamic distance possesses some remarkable properties, in particular its ability to automatize the search for low-energy reaction pathways and identify energetically metastable and stable states on the complex free energy surface. The dynamic distance, formulated within the general theory of reaction coordinates,[4] exhibits highly favorable mechanical and statistical properties which simplify the computation of free energies. In the following section we present the theory and general implementation of the RC. In sections 3 and 4 we demonstrate the versatility of the dynamic distance with applications employing both ab initio and classical molecular dynamics. As the choice of examples shows, while the dynamic distance is an extremely versatile constraint, it is particularly suited to the study of association and dissociation events and proton-transfer reactions, processes that play an extremely important functional role in biological systems.

2. Theory

Consider a system with 3N Cartesian coordinates or N position vectors in a configuration $\mathbf{x}$ given by $\mathbf{x} = (x_1....x_{3N}) = (\mathbf{r}_1....\mathbf{r}_N)$. The dynamic distance, D, defined as

$$D = \left(\sum_{\text{NOP}} \frac{\mu_{ij}}{\mu^*} (\mathbf{r}_i - \mathbf{r}_j)^2 \right)^{1/2} \qquad (1)$$

is the rms sum of distances between selected nonoverlapping pairs (NOP) of atoms, whose positions, $\mathbf{r}_i(t)$ and $\mathbf{r}_j(t)$, are time-dependent during the simulation. The square of each distance is weighted with the associated reduced mass, $\mu_{ij} = m_i m_j/(m_i + m_j)$, divided by an arbitrary constant mass. Setting μ^* to be the sum over all reduced masses, $\mu^* = \sum_{\text{NOP}} \mu_{ij}$, the dynamic distance, D, becomes the usual rms distance if the reduced masses of all atom pairs are identical. The reaction coordinate can be employed to drive a reaction by application of a time-dependent constraint, $D = D(t)$, or to relax the system and to sample characteristic quantities at intermediate positions using scleronomic constraints (D = const). The use of D as a restraint in umbrella sampling simulations will be discussed briefly at the end of this section.

We now consider the RC as a function $\tilde{D}(\mathbf{x})$ of the Cartesian coordinates (or position vectors) and the constraint $\sigma(\mathbf{x}) = \tilde{D}(\mathbf{x}) - D = 0$. For an atom i which belongs to one of the selected atom pairs, the constraint force is given by

$$\mathbf{f}_i^{\mathrm{c}} = \lambda \frac{\partial \tilde{D}}{\partial \mathbf{r}_i} = \frac{\lambda \mu_{ij}}{D\mu^*} (\mathbf{r}_i - \mathbf{r}_j) \qquad (2)$$

When the leapfrog algorithm is used to integrate Newton’s equations (or more precisely the Lagrange equations of the first kind)

$$\dot{\mathbf{r}}_i = \mathbf{v}_i \text{ and } \dot{\mathbf{v}}_i = \mathbf{F}_i/m_i + \mathbf{f}_i^{\mathrm{c}}/m_i \qquad (3)$$

the numerical form becomes

$$\mathbf{r}_i(t+\Delta t) = \underbrace{\mathbf{r}_i(t) + \Delta t\left(\mathbf{v}_i(t-\Delta t/2) + \Delta t \mathbf{F}_i(t-\Delta t)/m_i\right)}_{\mathbf{r}_i^*} + \Delta t^2 \mathbf{f}_i^c(t-\Delta t)/m_i = \mathbf{r}_i^* + \delta\mathbf{r}_i \qquad (4)$$

where $\mathbf{r}_i$* is the result of an unconstrained step in the time interval, Δt, under the influence of the force, $\mathbf{F}_i$. The Lagrange parameter $\lambda \equiv f^{\mathrm{c}}$, being the same for all atoms defined in the RC is usually called the ‘constraint force’ and is determined such that the constraint is satisfied. In this case, λ can be calculated directly from $\mathbf{r}_i$* and $\sigma(\mathbf{x})$ by means of a quadratic equation with no need for iteration. It can also be shown using (3) that there exists an analytical form for the constraint force

$$\lambda = -\frac{2K}{D} - \frac{1}{D} \sum_{\text{NOP}} (\mathbf{r}_i - \mathbf{r}_j) \cdot \frac{\mathbf{F}_i m_j - \mathbf{F}_j m_i}{m_i + m_j} \qquad (5)$$

where K represents the kinetic energy of the atoms involved in the constraint. This simple expression arises from the inclusion of the mass-weighting term and provides a numerical check for the constraint force calculated from the correction in (4).

Mechanical Properties. The dynamic distance formulated as described above possesses several favorable mechanical properties. First, as the RC is a function of interatomic distances which are internal coordinates, application of the constraint induces neither rotation nor translation of the system. Second, the mass-weighting procedure ensures the homogeneous action of the constraint across the system. This can be readily proven: Using the constraint forces (2) and the definition of the reduced mass, one finds that for the change of a distance due to the action of the constraint in lowest order

$$\Delta(\mathbf{r}_i - \mathbf{r}_j) = \frac{1}{2} (\Delta t)^2 \left(\frac{\mathbf{f}_i}{m_i} - \frac{\mathbf{f}_j}{m_j} \right) = \frac{1}{2} \text{const} \cdot (\mathbf{r}_i - \mathbf{r}_j) \qquad (6)$$

Obviously the relative change is the same for each atom pair, and, in particular, lighter atoms such as hydrogen are not influenced disproportionately by the constraint.

Statistical Properties. Phase space statistics can be determined by the mass-metric tensor which results in the so-called Fixman determinant[14]

$$z = \det(\mathbf{H}) = \sum \frac{1}{m_i} \left(\frac{\partial D}{\partial x_i} \right)^2 \qquad (7)$$

In the present case, $z = 1/\mu^*$, which is a constant and constitutes the statistical advantage of the dynamic distance. The immediate consequence proven by Fixman[14] is a coincidence of the probability density function (pdf) of the unconstrained system in configurational space, $P(q_i,D)$, and the pdf of the system constrained to constant D, $P_c(q_i;D)$, $\{q_i,D\}$ being a complete set of generalized coordinates. As recently shown,[11] the free energy can be obtained in a straightforward manner directly from the constraint force and the Fixman determinant. As the Fixman determinant is constant in this particular case, the relevant formula simplifies to

$$A(D) = \int \langle \lambda_D \rangle_c \mathrm{d}D \qquad (8)$$

and the free energy is simply the integral over the constraint force without any correction required. Equation 8 still holds when further coordinates, such as bond lengths are constrained as long as they do not interfere with the constraint on D.[15] For activated processes along an RC, Carter et al.[16] have derived an expression for the rate and have shown that the Fixman determinant determines the effective mass associated with the RC chosen. For the dynamic distance, the rate, k, for the escape from a stable state (around a minimum) between D_0 and $D^\ddagger$ over a barrier (at a maximum of free energy) at $D^\ddagger$ is

$$k = \kappa k^{\mathrm{TST}} = \kappa \sqrt{\frac{k_B T}{2\pi\mu^*}}\, P(D^\ddagger) / \int_{D_0}^{D^\ddagger} P(D)\mathrm{d}D \qquad (9)$$

where k^{TST} is the rate given by transition state theory (TST), κ is the transmission coefficient, and $P(D)$ is the one-dimensional pdf related to the free energy, $A(D)$, by

$$P(D) = \mathrm{const}\cdot\exp(-A(D)/k_B T) \qquad (10)$$

Apparently the statistical advantage of a constant Fixman determinant is essentially due to mass-weighting and has wide consequences stated in eqs 8 and 9. The expression 9 for the rate depends on the definition of the RC but not on the computation of free energy profiles from the constraint force according to eq 7. Therefore, eq 9 also holds for profiles calculated by umbrella sampling[8] or umbrella integration.[10,17] Forces derived from the umbrella restraint potential, $\sigma^2(\mathbf{x}) = (\tilde{D}(\mathbf{x}) - D)^2$, induce neither rotation nor translation when employing the dynamic distance.

3. Ab Initio Molecular Dynamics Applications

3a. Double Proton-Transfer Reaction in Formic Acid Dimer. The dynamic distance constraint is readily implemented within the framework of ab initio molecular dynamics. In this section we present the application of the dynamic distance constraint using Car−Parrinello molecular dynamics to study proton-transfer events and dissociation processes. We first apply the dynamic distance constraint to the study of the well-known double proton-transfer event (DPT) in the model compound formic acid dimer shown in Figure 1. For this simple example, we discuss the technical details concerning the implementation of the constraint and show how one can extract accurate free energy profiles.

Methods. All calculations were performed using the CPMD 3.4 package.[18] The formic acid dimer was placed in a periodically repeating cell with dimensions 13.25 × 13.25 × 13.25 Å^3. A fictitious mass of 400 au was ascribed to the electronic degrees of freedom within the Car−Parrinello scheme. The coupled equations of motion for atomic nuclei and molecular orbitals were solved using the velocity Verlet algorithm with a time-step of 4 au. For each nuclear configuration, the Kohn−Sham equations were solved using the BLYP exchange-correlation functional.[19,20] Core electrons were treated using norm-conserving Troullier-Martins pseudopotentials,[21] and the valence electrons were expanded in a plane wave basis up to an energy cutoff of 70 Ry in all simulations performed.

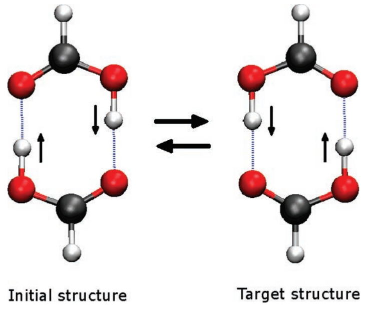


Figure 1. Illustration of initial and target structures for the double proton exchange in the formic acid dimer.

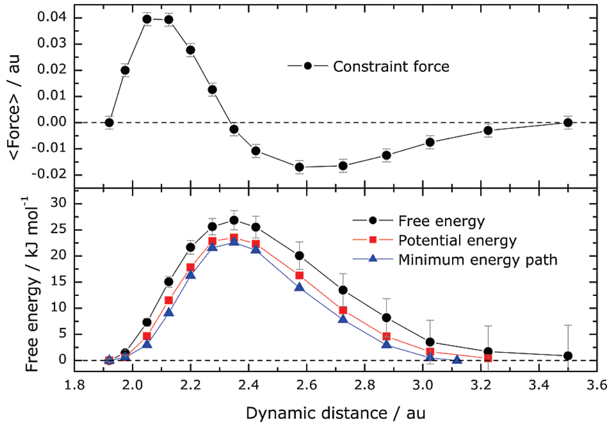


Figure 2. Average constraint force (top) and energy profiles (bottom) along the reaction coordinate for formic acid dimer. The free energy curve (black) was obtained by integration of the force curve (top). The average finite temperature potential energy is shown in red, the minimum energy profile in blue. Energies are given relative to their starting values.

The unconstrained system was first brought to thermodynamic equilibrium at 300 K using a Nosé-Hoover thermostat.[22] For the constrained CP−MD simulations, the dynamic distance constraint comprises two distances which represent the two O−H chemical bonds. The dynamic distance was initially set at a value of 1.925 au, which was the average value of the dynamic distance in a 0.5 ps unconstrained MD simulation, and subsequently was systematically increased. The chosen increment in the step size was very small in the initial stages of the reaction during the cleavage of the O−H chemical bonds. For each dynamic distance, D_i, the system was re-equilibrated before starting a 1 ps 'production run'. The simulation length employed provides sufficiently reliable average constraint forces. Free energies were calculated by numerical integration from the cumulative average of the constraint forces using eq 8. The entropy contribution to the free energy profile was calculated from the eigenvalues of the mass-weighted covariance matrix for each constrained CP−MD simulation.[23,24]

Results. Figure 2 shows the average constraint force and free energy profile along the reaction coordinate for formic acid dimer. We observe the well-known concerted double proton-transfer event. The average constraint force starts at zero and rises to a maximum at 2.1 au. This positive constraint force arises from the fact that the constraint is driving the system away from the stable configuration, as it 'pushes' the protons across the hydrogen bonds, causing the O−H chemical bonds to break. The average constraint force then falls to zero at 2.35 au, which defines the transition state for the reaction. At larger dynamic distances the average constraint force becomes negative, which represents the constraint acting to 'hold back' the protons as they try to complete the DPT reaction. The concerted nature of the reaction is represented by the single energy barrier in the free energy profile, which reaches a maximum of 26.7 kJ/mol at the transition state (a dynamic distance of 2.35 au). Quasi harmonic frequencies ω_i were calculated from the mass-weighted covariance matrix and inserted into the entropy formula[23,24]

$$S_{ho} = \sum_{i=1}^{3N-6} k_B(\hbar\omega_i/k_BT)(\exp(\hbar\omega_i/k_BT) - 1) - \ln(1 - \exp(-\hbar\omega_i/k_BT)) \quad (11)$$

After subtracting the entropic contribution, the resulting enthalpy profile for the reaction is very similar to the minimum energy path (MEP), as shown in Figure 2. The small differences arise from the fact that the MD simulation at 300 K is probing more configurational space than the MEP. This is most noticeable for larger dynamic distances, where the constraint is now acting on the newly formed hydrogen bonds. The free energy and enthalpy profiles presented in Figure 2 clearly demonstrate how well the dynamic distance constraint controls the DPT reaction, in comparison to previous constraints.[25]

To obtain correct free energy profiles, it is necessary to verify that the simulation length for each constrained CP−MD simulation provides a sufficiently reliable average constraint force. This is best achieved by monitoring the cumulative average constraint force.

Figure 3 shows the variation in the constraint force and cumulative average constraint force over a 1 ps CP−MD simulation at a dynamic distance constraint of 2.125 au. While the constraint force varies quite significantly across the trajectory from 0.01 to 0.065 au, the cumulative average constraint force converges within 1 ps to an average value of 0.039 au. The magnitude of the fluctuations in the cumulative average constraint force over the last 400 fs of the trajectory provides an estimate of the error used when calculating the free energy profile as shown in Figure 2. The rate of convergence of the cumulative average constraint force is system specific and must therefore be determined for the particular system of interest. In the case of formic acid dimer, as shown in Figure 3 a 1 ps trajectory is sufficiently long to liberate accurate converged average constraint forces.

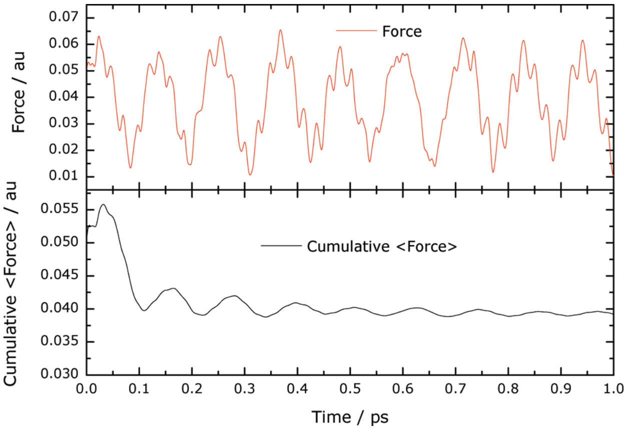


Figure 3. Trajectory of the constraint force (top) and cumulative average (bottom) as obtained in a 1 ps run at a constant reaction coordinate, $D = 2.125$ au for FAD.

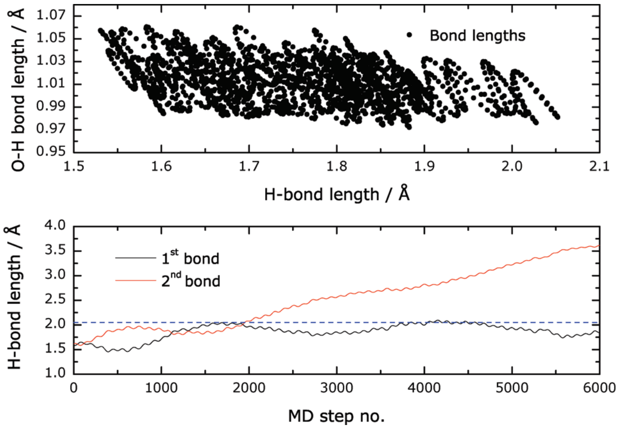


Figure 4. The lengths of a O−H chemical bond and an H-bond over a 0.5 ps unconstrained CP−MD run (top). The H-bond lengths in a constrained CP−MD run (bottom) demonstrate that one bond breaks after 2000 steps.

3b. Dissociation of Formic Acid Dimer. We extended our study of formic acid dimer to investigate the dissociation of the dimer using the dynamic distance constraint. The principal aim of this analysis was to determine whether the dissociation process is concerted or stepwise. In order to look at this event, we implemented the constraint in a slightly different way: Instead of performing a series of constrained CP−MD simulations, each at a specific constant dynamic distance, we started a simulation at a dynamic distance value of 3.025 au and systematically increased the dynamic distance by 0.0004 au per step across a single trajectory. In this case, the dynamic distance constraint comprised the two hydrogen bonds. The systematic increase in the constraint therefore drives dissociation of the dimer. The growth rate of 0.0004 au per step is small enough that the kinetic energy of the electrons remains unperturbed during the simulation. In the lower panel of Figure 4, we show the observed change in the two hydrogen bond lengths across the trajectory. Initially, both hydrogen bond lengths have a value of approximately 1.6 Å. On increasing the dynamic distance constraint over the first 2000 steps, both hydrogen bond

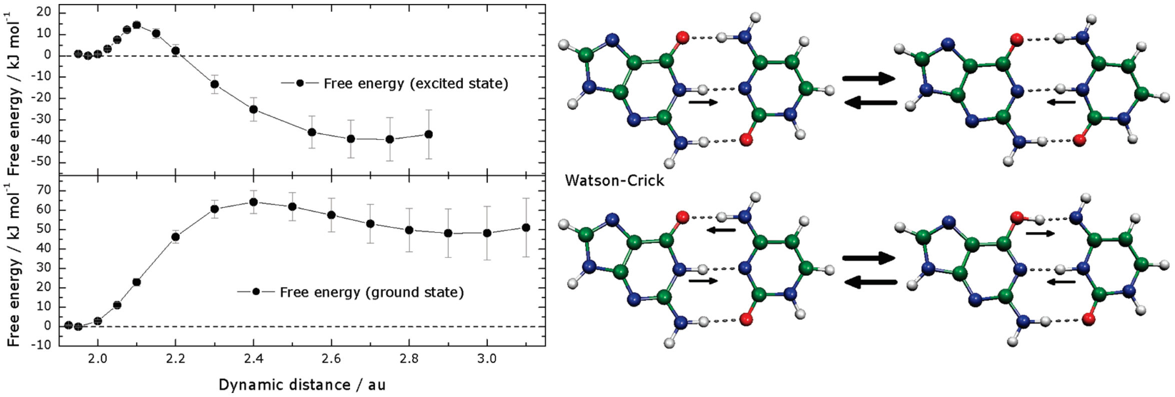


Figure 5. Free energy profiles for the ground (lower panel) and excited (upper panel) state proton-transfer reactions in the G−C base pair. The reactant and product states for the two reactions are shown graphically on the right.

lengths increase to approximately 2.0 Å. After this point, one of the hydrogen bonds breaks, and the associated hydrogen bond length increases to over 3.5 Å. The other hydrogen bond fluctuates, and its associated hydrogen bond length varies between 1.8 and 2.05 Å. The upper panel of Figure 4 shows the variation in the OH···O hydrogen bond length and the O−H chemical bond length over a short 0.5 ps unconstrained CP−MD simulation at 300 K. The hydrogen bond length varies up to 2.05 Å. These results clearly demonstrate that the dissociation process occurs in a stepwise fashion, a result that is consistent with that observed in CP−MD simulations at higher temperatures (results not shown). This simple example illustrates the versatility of the dynamic distance constraint, which can be used to probe both the DPT event and the dimer dissociation process. It also underlines the fact that the constraint does not favor or bias the reaction mechanism, be it concerted (DPT) or stepwise (dissociation).

3c. Guanine−Cytosine DNA Base Pair: Ground-State Proton Transfer and Excited-State Coupled Proton−Electron Transfer. A further interesting aspect of the dynamic distance reaction coordinate concerns its remarkable predictive properties: Unlike atom-specific, local constraints, such as simple distance and angle constraints, which are chosen in advance in order to drive a system to a known predefined product state, the dynamic distance is a flexible collective reaction coordinate that comprises multiple internal degrees of freedom. As such, when implemented appropriately, the dynamic distance constraint automatizes the search for the lowest energy reaction pathway(s) without any specific a priori knowledge of the product state. A good example of the predictive properties of the dynamic distance constraint can be found in a recent study of irradiation-induced damage mechanisms in the guanine−cytosine (G−C) base pair,[26] and the reader is referred to this reference for computational details. For the purposes of this paper, we merely summarize the general result.

The G−C base pair possesses three interbase hydrogen bonds. Starting in the Watson−Crick geometry, there exist a large number of possible single or multiple proton-transfer reactions that can bring the system to a variety of different hydrogen-bonded tautomeric states. The dynamic distance constraint for this system was constructed using the three N−H chemical bonds involved in interbase hydrogen bonding. Similar to the method described in section 3a, we performed a series of constrained Car−Parrinello MD simulations in order to identify the lowest energy PT reactions in both the ground and the excited electronic state of the G−C base pair. The results are summarized in Figure 5. In the ground state, we observe a double proton-transfer event over a large free energy barrier (64.3 kJ/mol) leading to a meta-stable product state. In contrast to this, for the singlet excited state, we observe a single proton-transfer event over a very small free energy barrier (14.3 kJ/mol) leading to an energetically favorable charge-transfer product state. It is important to recognize that the same constraint, implemented in exactly the same manner produces two completely different reaction pathways, because the underlying potential energy surface which is calculated ‘on-the-fly’ during the constrained CP−MD simulations is different in the ground and excited states. For both reactions, the dynamic distance constraint finds the energetically most favorable reaction pathway, as confirmed by static post Hartree−Fock calculations.[27−29]

In these simulations, the constraint was defined using the three N−H chemical bonds for the interatomic distances. However, if the constraint is formulated with the three interbase hydrogen bonds, the average constraint force and free energy profiles look rather different: In the initial stages of the reaction the average constraint force profile rises gradually to a maximum as the system is driven toward the transition state. At the transition state, the average constraint force decreases directly to zero, and the new chemical bond(s) are formed immediately. The constraint loses control of the reaction as the flexible hydrogen bonds readily alter their geometry slightly to allow the chemical bonds to form directly while still fulfilling the conditions of the constraint. Similar behavior was also observed in the case of targeted molecular dynamics.[25] The resulting free energy barrier for the reaction is still accurately obtained by integration of the average constraint force as the constraint controls the reaction up to the transition state. Nevertheless, this simple example demonstrates that one must choose the specific interatomic distances carefully for the particular system in question in order to define a constraint that can control the reaction along the entire pathway.

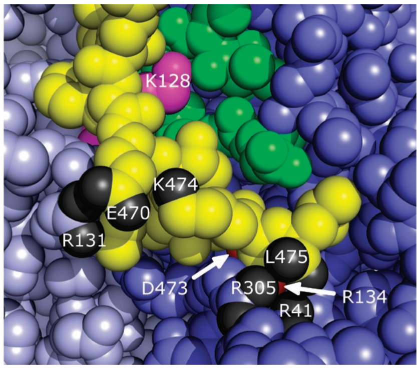


Figure 6. Three structural elements are sealing the active site of rubisco: C-terminal strand (yellow) with terminus L475, K128 (magenta) and loop 6 (green). Four ionic contacts are stabilizing the closed, active conformation of the C-terminal tail: The contacts E470-R131, L475-R41, and L475-R305 which are exposed to the solvent, and the buried bridge D473-R134 which is conserved in all rubisco homologues. The intrastrand salt bridge E470-K474 is not considered here.

4. Classical MD: Opening the Binding Pocket of Rubisco

The binding niche of rubisco is sealed by three structural elements of which the large subunit's C-terminal strand is the outermost. In the closed conformation this element stabilizes the catalytically active state of the protein (see Figure 6).

The composition of the C-terminal tail influences the substrate specificity of rubisco, which catalyzes the fixation of carbon dioxide and molecular oxygen. The time window hypothesis[30] ascribes this to the dynamics of the tail which may transiently lift off, thus interrupting catalysis. The C-terminal strand is attached to the underlying protein corpus by several ionic bridges,[31] whose specific number varies from homologue to homologue. In order to analyze the structural dynamics of the C-terminal tail, these contacts were cleaved using the dynamic distance constraint in the framework of classical MD simulation. Free energy profiles were calculated, and the contribution of each salt bridge to the stability of the enzyme's closed, active conformational state was estimated.

Methods. The four salt bridges are all of the following type: $-C-O_2^{-}\cdots H_2^{+}-N-C^{\zeta}-$ (arginine). In order to allow dynamical exchange among the carboxyl oxygens or guanine hydrogens and to avoid interference with the bond length constraints imposed on the NH_2^+ moiety, the dynamic distance of eq 1 is defined here as the rms carbon−carbon distance for each of the four salt bridges. For the classical MD simulations, the GROMACS simulation package[32] was employed with explicit SPC water and the force-field 43A1. The simulations were performed on the rubisco structure (1RBL) from the reference organism Synechococcus sp. PCC6301. The two large subunits forming a functional L2 protomer including two binding niches were treated explicitly, while the missing adjacent subunits were emulated by restraining harmonic potentials ($k_{xyz} = 250$ kJ mol^{-1} nm^{-2}) on heavy atoms involved in polar contacts to the neighboring subunits. The binding niche was modeled, and the substrate RuBP and the carbamylated K201 were parametrized as described previously.[31] Crystal waters were retained, and the protein was inserted into a simulation cell flooded with bulk water. The resulting system contained a total of 78 854 atoms. After an initial energy minimization, the system was heated and brought to equilibration during a short 200 ps MD simulation. A suitable electrostatic cutoff and reaction field were used ($r_{cp} = 0.8$ nm, $r_{cl} = 1.4$ nm, $r_{crf} = 1.4$ nm, $\epsilon_{rf} = 54$). Bonds involving hydrogen atoms were constrained using LINCS, and a time-step of 1 fs was employed. All runs were performed at a temperature of 298.15 K and a pressure of 1 bar, both regulated by a Berendsen thermostat and barostat, respectively. After equilibration, a number of conformations were extracted every 1000 ps from an unconstrained simulation to be used as starting conformations for the constrained runs.

Pathways and Free Energy. A pathway can be generated by modulating the RC from an initial to a final value during a so-called slow-growth simulation run. However, the computation of the free energy profile (instead of the work profile as immediately yielded by the slow growth run) requires converged mean constraint forces obtained at discrete points along the reaction pathway during a series of relaxation runs.[2,33] Due to the their rugged and heavily structured energy landscape,[34] difficulty arises when calculating free energy profiles for protein systems: In each relaxation run, the system may evade into different pathways thus rendering structurally discontinuous trajectories and useless free energy profiles.[35] To prevent such incidents a novel variant of the equidistant relaxation protocol was implemented in the rubisco simulations. The stop-and-go-like (SNG) approach integrates the slow-growth and the relaxation phases into a single simulation; the system is equilibrated for a certain period of time in a scleronomic "stop" phase ($D =$ const) in which the average constraint force is calculated. The transition is then driven further in a rheonomic "go" phase ($dD/dt > 0$). This procedure is repeated for a certain number of equidistant points on the RC. The average constraint forces of the stop phases are then integrated to obtain the free energy profile of the particular reaction path.

Optimization. Before performing the production runs, it is necessary to optimize several parameters for the SNG approach, paying consideration to the available computational resources. These parameters include the number of equidistant relaxation points, p, on the RC, the equilibration time allowed at each of these points (relaxation phase: t_{stop}), and the fraction of this time period used to calculate the average constraint force (measuring phase: $t_{av} \leq t_{stop}$). The total resulting simulation time is $t_{stop} + (p - 1)(t_{go} + t_{stop})$. The RC was increased in steps of 1 nm from a starting value 0.45 nm, and the final constraint value was inferred from unconstrained long-term simulations (5−10 ns) in which the four salt bridges were observed to rupture spontaneously.

It was found that a relatively long relaxation period was required in order to obtain well-converged average constraint forces; however, the large forces observed in the initial stages of the relaxation can safely be discarded. Optimum convergence of the constraint force was achieved when considering

Table 1. Results of the Parameter Optimization in Three *ceteris paribus* Parameter Groups[a]

p	t_{tot} (fs)	t_{go} (fs)	t_{stop} (fs)	t_{av} (fs)	$\langle df^{c}_{cum}/dt \rangle$ (kJ mol^{-1} nm^{-1} ps^{-1})	$\langle \sigma(\langle f^{c} \rangle) \rangle$ (kJ mol^{-1} nm^{-1})
20	100019	1	5000	2500	**−20.48**	**9.44**
200	100199	1	500	250	−181.67	133.26
2000	101999	1	50	25	1029.91	253.88
20000	99999	1	4	2	−5086.69	85.49
20	100019	1	50000	25000	**−1.04**	**9.75**
200	100199	1	5000	2500	−2.84	32.09
2000	101999	1	500	250	−30.28	98.53
20000	99999	1	49	25	−34.58	255.08
20	10140	260	260	130	−75.31	138.29
20	101400	2600	2600	1300	−16.99	41.38
20	1014000	26000	26000	13000	**−0.42**	**14.37**

[a] Maximal distribution of simulation time on the scleronomic phases (~0.1 ns and ~1 ns, respectively, number of points variable), and equal distribution to the scleronomic and rheonomic phases (20 discrete points, simulation time variable). Optimal performance values are printed in boldface. 1 ns runs with an equal distribution of simulation time for stop and go phases yields the best results (last line).

only the constraint forces obtained during the latter half of the scleronomic phases ($t_{av} = 0.5 t_{stop}$).

The following performance parameters were considered for further optimization: $\langle df^{c}_{cum}/dt \rangle$: The slope of the cumulative average of the constraint force f^{c} averaged during the relaxation period of length t_{av} as a criterion for the convergence of the constraint force. $\langle \sigma(\langle f^{c} \rangle) \rangle$: The standard error of the average constraint force averaged during the relaxation period as a criterion for the quality of the mean constraint force.

Table 1 shows the average values of these quantities for each run. The optimal parameter set was derived by the *ceteris paribus* principle, i.e., each independent variable was changed while keeping all others fixed. As optimal convergence of the constraint force at each discrete point along the reaction coordinate is the primary objective, the first approach taken was to maximize that portion of the simulation time spent on the scleronomic phases (relaxation). As a consequence, in these runs only 1 fs was spent driving the system in each rheonomic phase. The number of discrete points was varied as well as the length of each scleronomic phase accordingly. This was done in two *ceteris paribus* groups for total simulation times of ~0.1 ns and ~1.0 ns. In the third group, the available simulation time was equally distributed between the stop and go phases which yielded considerably better results. The total simulation time was set to ~0.01 ns, ~0.1 ns, and ~1 ns. For this group, the cumulative constraint force convergence is shown in Figure 7.

The results summarized in Table 1 show that in the present case, an equal distribution of simulation time between the rheonomic and scleronomic phases provides the best results in terms of the criteria defined above. While it appears sufficient to calculate the average constraint force at only a small number of points along the reaction coordinate, slow reaction coordinate modulation and long time scale equilibration phases are essential for good results. As a rule of thumb, it is usually sufficient to check convergence of the constraint force for the starting structure only to get an impression of the necessary duration of the scleronomic phase.

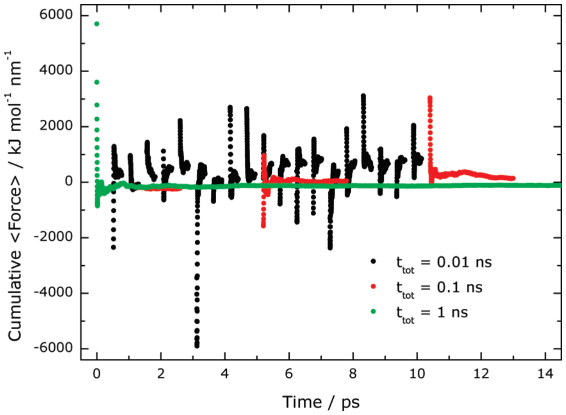


Figure 7. Convergence of the cumulative constraint force at various total simulation times (0.01, 0.1, and 1 ns). Only a sufficiently long simulation time allows convergence of the constraint force during the relaxation phase.

In consideration of the general problem of accuracy in free energy calculations,[33] the convergence of mean forces was checked in all cases. Extensive studies on the protein showed that 1 ns runs produce reliable profiles without discontinuities with the appropriate time allocations for driving, equilibrating, and averaging. This is important because the back reaction—often recommended as a test—cannot be simulated to the same degree of accuracy for complex activated processes in large systems.

Results. Four 1 ns production runs on rubisco were performed with the optimized parameters described above. Although the free energy profiles of all simulations were consistent in terms of convergence and error of the constraint force, the specific forms of the free energy profiles were somewhat different (data not shown). We conclude that the system takes different pathways depending on the specific initial geometry. Nevertheless the sequence of rupturing events along the reaction coordinate and the relative contribution of each salt bridge to the stability of the enzyme's active conformational state were the same in all runs. A representative example is shown in Figure 8. The carboxy-terminal contacts of L475 are seen to open first; these are easily solvated and only play a minor role in stabilizing the closed conformational state at room temperature. The highly conserved bridge between D473 and R134 is cleaved last and thus is the main player, while the contact E470-R131 plays a modulating role, which makes it sensitive to specificity enhancing mutations. These results are in good agreement with previously published data based on a combined bioinformatic tools and TMD simulation approach.[31]

For the reaction path depicted in Figure 8 we calculated a free energy barrier of $\Delta A = 25.37 \pm 3.84$ kJ/mol. The error was estimated as described previously.[2] Figure 9 shows the underlying profile of the mean constraint force. Similar values of ΔA were obtained from other production run pathways, though this does not exclude the possible existence of further reaction pathways with lower activation barriers.

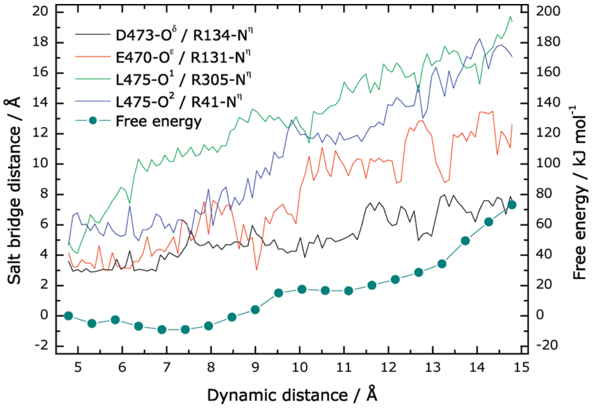


Figure 8. Free energy profile and associated distances of the four salt bridges as a function of the dynamic distance constraint. The conserved salt bridge D473/R134 is the last to open. The carboxyl-terminal contacts of L475 have a negligible effect, while the E470/R131 bridge plays a modulating role in the stability of the closed C-terminal strand.

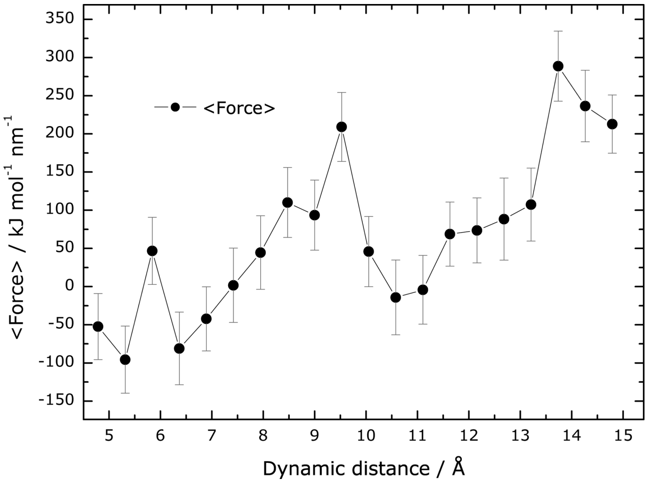


Figure 9. Profile of the average constraint force with error bars across the reaction coordinate.

To a single order of magnitude, the energy barrier is in good agreement with the nanosecond time scale of the C-terminal strand's opening in the long-term simulations already mentioned, though experimental rates are not available yet.

The path taken by the C-terminal strand during the enforced transition was confirmed by comparing an intermediate structure to results obtained from the long-term free MD simulations mentioned above. Figure 10 compares snapshots obtained from the constrained and free MD simulations at the moment when the crucial salt bridge D473/R134 is ruptured. All elements of the fluctuating system are seen to adopt comparable configurations.

5. Conclusions

In this paper we have introduced a novel versatile reaction coordinate, the 'dynamic distance', which has been specifically designed for the study of reactions and activated processes involving the cleavage and/or formation of a set of bonds or contacts. The flexible reaction coordinate, formulated as a mass-weighted mean of selected distances does not bias or favor the sequence or mechanism of events on the resulting reaction pathway. Due to the presence of the mass-weighting term, the free energy profile for the process is readily obtained by integration over the mean constraint force without any correction, and the rheonomic constraint driving this RC represents a minimal perturbation in that it causes no net momentum or torque.

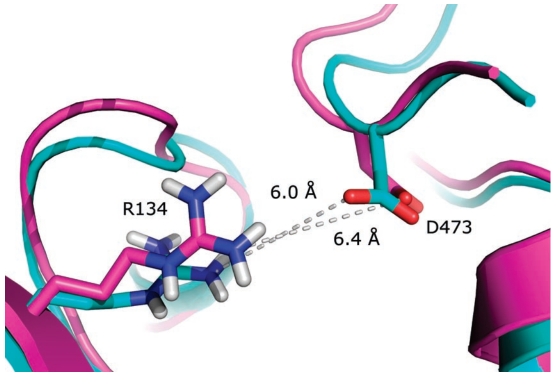


Figure 10. Comparison of structures with open C-terminal tail obtained from constrained (1 ns; dark) and free (5 ns; light) MD simulations. Both conformations are very similar, demonstrating the use of the constraint to induce accurate transitions in short simulation times.

Using several examples in the framework of both ab initio and classical molecular dynamics, we have demonstrated the versatility of the dynamic distance constraint, paying particular attention to both the implementation and optimization of the RC in order to obtain accurate reaction pathways and free energy barriers. Our study of the DPT event in the prototypical model system formic acid dimer reproduces the well-known concerted reaction mechanism, and the size of the activation barrier is in full agreement with previous studies. However, the free energy profiles clearly demonstrate an increased level of control compared to alternative constraints.[25] The predictive properties of the dynamic distance have been highlighted using a recent application of the constraint to automatize the search for the lowest energy proton-transfer events in the G−C base pair in both the ground and excited state.[26] Using classical MD simulation, we considered the opening of a binding pocket in a large protein−water system (~10^5 atoms). Proteins are known for their complex glasslike energy landscape[34] which opens a manifold of pathways for such complex processes. The constrained simulations produced both unproductive and productive pathways. The latter exhibit activation barriers and intermediate structures which compare well with available long time-scale free MD simulations. A common feature of all the simulations is the unique sequence of events when the crucial salt bridges are cleaved, which was in the focus of the present study.

Except for simple cases, there is no way to decide whether the best path was already detected by any method whatsoever. A second caveat concerns the directionality of coordinate driving methods which tend to produce different pathways during decreasing and increasing the reaction coordinate. Therefore, repeated simulations with different starting conditions and directions are suggested wherever possible for determining realistic pathways.

Among the broad range of potential applications, the dynamic distance reaction coordinate is particularly suitable for the study of important functional processes in biological systems involving association and dissociation events and proton-transfer reactions.

Acknowledgment. C.B. and J.S. acknowledge funding by the VolkswagenStiftung. One of the authors (N.L.D.) acknowledges funding by the German Science Foundation DFG with FOR 618. Another author (P.R.L.M.) acknowledges funding by the CEA. The authors are grateful to BOVILAB@RUB, EMBL-Heidelberg, NIC Jülich, and ZAIK Köln for computer time.

References

(1) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514−1527.

(2) Swegat, W.; Schlitter, J.; Kruger, P.; Wollmer, A. *Biophys. J.* **2003**, *84*, 1493−1506.

(3) Akola, J.; Jones, R. O. *J. Phys. Chem. B* **2003**, *107*, 11774−11783.

(4) Schlitter, J.; Swegat, W.; Mulders, T. *J. Mol. Model.* **2001**, *7*, 171−177.

(5) Davies, J. E.; Doltsinis, N. L.; Kirby, A. J.; Roussev, C. D.; Sprik, M. *J. Am. Chem. Soc.* **2002**, *124*, 6594−6599.

(6) Blumberger, J.; Sprik, M. *Theor. Chem. Acc.* **2006**, *115*, 113−126.

(7) Muller, R. P.; Warshel, A. *J. Phys. Chem.* **1995**, *99*, 17516−17524.

(8) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275−282.

(9) Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. *Mol. Simul.* **1993**, *10*, 291−308.

(10) Kästner, J.; Thiel, W. *J. Chem. Phys* **2005**, *123*.

(11) Schlitter, J.; Klähn, M. *J. Chem. Phys.* **2003**, *118*, 2057−2060.

(12) Ma, J. P.; Sigler, P. B.; Xu, Z. H.; Karplus, M. *J. Mol. Biol.* **2000**, *302*, 303−313.

(13) Ensing, B.; De Vivo, M.; Liu, Z. W.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73−81.

(14) Fixman, M. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 3050−3053.

(15) Schlitter, J.; Klähn, M. *Mol. Phys.* **2003**, *101*, 3439−3443.

(16) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472−477.

(17) Kästner, J.; Thiel, W. *J. Chem. Phys* **2006**, *124*.

(18) Hutter, J.; Ballone, P.; Bernasconi, M.; Focher, P.; Fois, E.; Goedecker, S.; Marx, D.; Parrinello, M.; Tuckerman, M. *MPI für Festkörperforschung and IBM Zürich Research Laboratory*; Stuttgart, 2001.

(19) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

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

(21) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993−2006.

(22) Nosé, S. *J. Chem. Phys* **1984**, *81*, 511−519.

(23) Schlitter, J. *Chem. Phys. Lett* **1993**, *215*, 617−621.

(24) Andricioaei, I.; Karplus, M. *J. Chem. Phys* **2001**, *115*, 6289−6292.

(25) Markwick, P. R. L.; Doltsinis, N. L.; Marx, D. *J. Chem. Phys* **2005**, *122*, 054112.

(26) Markwick, P. R. L.; Doltsinis, N. L.; Schlitter, J. *J. Chem. Phys.* **2007**, *126*, 45104−45107.

(27) Florian, J.; Leszczynski, J. *J. Am. Chem. Soc.* **1996**, *118*, 3010−3017.

(28) Sobolewski, A. L.; Domcke, W. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2763−2771.

(29) Sobolewski, A. L.; Domcke, W.; Hattig, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17903−17906.

(30) Schlitter, J.; Wildner, G. F. *Photosynth. Res.* **2000**, *65*, 7−13.

(31) Burisch, C.; Wildner, G. F.; Schlitter, J. *FEBS Lett.* **2007**, *581*, 741−748.

(32) van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. *Gromacs User Manual version 3.2*; 2004. www.gromacs.org (accessed March 2004).

(33) Mark, A. E.; van Helden, S. P.; Smith, P. E.; Janssen, L. H. M.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **1994**, *116*, 6293−6302.

(34) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598−1603.

(35) Klähn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 15645−15650.

CT700170T

JCTC Journal of Chemical Theory and Computation

Biased Molecular Simulations for Free-Energy Mapping: A Comparison on the KcsA Channel as a Test Case

Enrico Piccinini,*,[†,‡] Matteo Ceccarelli,[§] Fabio Affinito,[†,||,⊥] Rossella Brunetti,[†,||] and
Carlo Jacoboni[†,||]

CNR-INFM National Research Center on nanoStructures and bioSystems at Surfaces (S^3), Via Campi 213/A, I-41100 Modena, Italy, Dipartimento di Ingegneria Elettronica, Informatica e Sistemistica DEIS, Alma Mater Studiorum Università di Bologna, Viale Risorgimento 2, I-40136 Bologna, Italy, Dipartimento di Fisica and Sardinian Laboratory for Computational Materials Science - SLACS, Università di Cagliari, Cittadella Monserrato, I-09042 Monserrato (CA), Italy, and Dipartimento di Fisica, Università di Modena e Reggio Emilia, Via Campi 213/A, I-41100 Modena, Italy

Received July 31, 2007

Abstract: The calculation of free-energy landscapes in proteins is a challenge for modern numerical simulations. As to the case of potassium ion channels is concerned, it is particularly interesting because of the nanometric dimensions of the selectivity filter, where the complex electrostatics is highly relevant. The present study aims at comparing three different techniques used to bias molecular dynamics simulations, namely Umbrella Sampling, Steered Molecular Dynamics, and Metadynamics, never applied all together in the past to the same channel protein. Our test case is represented by potassium ions permeating the selectivity filter of the KcsA channel.

1. Introduction

Molecular Dynamics (MD) simulation is considered today the most powerful computational method to explore or interpret specific protein functions, provided that a high-resolution structure is known, and it has been widely applied to study specific features of single-ion translocations through nanometric membrane channels that underlie many important physiological features.[1]

The power of the method relies on the possibility of linking specific features of the permeation path with the peculiar interactions existing among the permeating ions and between each of them and the protein residues facing the permeation pathway.

The main limitations of the method are due to (a) the parametrization of the force field and its accuracy to take into account the strong electrostatic interaction between ions and proteins, (b) the computationally expensive very large number of atoms forming the simulated system, and (c) the way the complex electrostatics of the nanometric environment is tackled.

With reference to point (a) above, Allen et al. recently compared the most widely used biomolecular force fields using gramicidin A as a prototypical narrow ion channel showing that a polarizable reliable force field would introduce a significant enhancement of state-of-the-art results.[2]

The second problem listed above prevents the possibility to directly compare results from MD simulations with experimental results. It is in fact known that single-ion translocations require typical times ranging from 10 to 100 ns. With the present hardware and software computation tools only a few events can be observed with MD: they are useful

* Corresponding author phone: +39 059 205 5292; fax: +39 059 205 5616; e-mail: enrico.piccinini@unimore.it. Corresponding author address: CNR-INFM National Research Center on nanoStructures and bioSystems at Surfaces (S^3), Via Campi 213/A, 41100 Modena, Italy.

[†] CNR-INFM National Research Center on nanoStructures and bioSystems at Surfaces (S^3).

[‡] Università di Bologna.

[§] Università di Cagliari.

[||] Università di Modena e Reggio Emilia.

[⊥] Current address: International School of Advanced Studied (SISSA/ISAS), Via Beirut 2/4, 34014 Trieste, Italy.

10.1021/ct7001896 CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/11/2007

to study the full permeation pathway but still not enough for the straightforward simulation of a macroscopic ion flux. To fill this gap computational approaches able to calculate ion fluxes and including as much as possible the molecular information of the protein in the input parameters and in the model have been recently presented in the literature.[3−5] They all rely on more or less detailed information about the potential of mean force (PMF) of the system formed by the ions and the protein to identify the relevant occupation configurations involved during the permeation process and the probabilities associated with the transitions between them. The use of these "mesoscale" simulation procedures allows the linking of the atomistic description to the real functional properties of the proteins and promises to become in the future one of the investigation tools to be used for engineering protein functionality and fixing failures.

Mapping all the relevant structures of a complex PMF from MD simulations is a very difficult task, especially in view of the fact that an uncertainty of few k_BT in the evaluation of a free-energy barrier can be very relevant when the barrier height is used to estimate the transition probabilities in the simulation of the conduction process.[5] An accurate MD estimate can be obtained with the use of many reaction coordinates and long computer runs, which in some cases makes the calculation in practice impossible. Thus very relevant care is devoted to study when and how the simulation problem can be suitably simplified to find a reasonable trade-off between accuracy and resource demand.

For this purpose many computational techniques have been proposed and applied in the literature to artificially "bias" a simulation and force the time evolution of the system toward a given transition of interest, depending on the case at hand. Among them, we have chosen three methodologies relying on very different computational strategies, namely Steered MD[6] (SMD), the most widely used Umbrella Sampling[7] (US), and Metadynamics[8] (MetaD).

The main aim of this investigation is to establish the degree of reliability and the vulnerable aspects of these computational techniques on the basis of a common test on a nanometric channel. To this purpose we report both the evaluations of the free-energy profiles and the corresponding technique-dependent error. This critical analysis seldom accompanies this kind of calculations. Our test case is potassium ions permeating the selectivity filter of the bacterial potassium channel KcsA from *Streptomices lividans*.[9]

The KcsA structure is known from X-ray investigations, further confirmed by MD simulations. The potassium permeation of this channel takes place through a short and narrow region of the protein, called the "selectivity filter''. Seven stable binding sites have been identified, usually referred to as S_{ext}, S_0, ..., S_4, S_{cav} in which, alternatively, potassium ions and water molecules are found. The two outermost sites S_{ext} and S_0 were both first predicted in MD simulations[10] and subsequently observed in the higher resolution X-ray structure.[9] S_{ext} can fictitiously be made collapsing onto S_0 when the conduction properties of the channel are investigated because its position is diffuse and quite close to the bulk water phase. Two ions must always reside in the region S_1, ..., S_4 in a stable conductive situation, otherwise the protein changes its conformation and switches to a nonconductive state. The conduction process involves the simultaneous and concerted movement of ions in a single file, giving origin to a cycle of different occupancy configurations, as indicated in the sketch reported in Figure 1 (site S_{ext} is not represented). A proper free-energy barrier identifies each transition between different configurations. In this study we are interested in mapping the free-energy profile associated with internal transitions (i.e., transitions not involving new ion entries or exits), labeled as (t) or (d) in Figure 1.

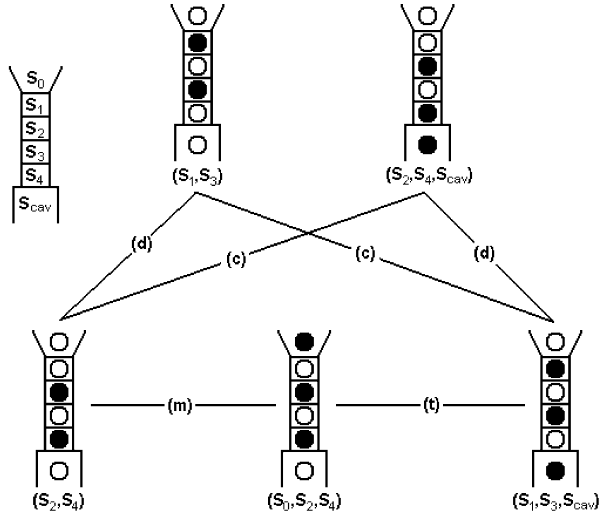


Figure 1. Configurations considered in the model and transitions among them. A sketch of the selectivity filter is reported on the left; site S_{ext} is located on top of site S_0, and it is not represented. Open circles stand for water molecules, and solid circles stand for potassium ions. Labels (m), (c), (d), and (t) mean an ion entry/exit into/from site S_0, for an exit/entry from/into the cavity site S_{cav}, for a two-ion concerted motion, and for a three-ion concerted motion, respectively.

The choice of the KcsA channel as the test case for our analysis is justified by the fact that in this highly selective nanometric channel ions move in single file along a pore which is roughly their size and strongly interact with each other and with the protein environment, thus producing physical challenging conditions for any molecular simulation.

Moreover, this system was deeply investigated in the past by means of the US technique, and many results to compare it with can be found in the literature.[3,10−12] To our knowledge MetaD was exploited in the past to study chlorine channels,[13] and it is here applied to the KcsA case for the first time.

2. Methods

2.1. Modeled System. Our starting point is the most recent X-ray structure of the KcsA solved at 2.0 Å resolution (PDB code 1K4C), inserted in a slab composed by 500 octane molecules mimicking the cell membrane. We solvated with 8802 water molecules, and 24 chlorine ions were used to keep the system electrically neutral. The final system, analogous to the one reported in the literature,[14] is composed of 34 434 atoms.

Calculations have been carried out with the GROMACS 3.3 package[15,16] for SMD and US simulations and with the ORAC code[17] for MetaD, using in all cases the standard GROMACS force field (also known as GROMOS 87) for the protein and SPC model for water. This combination of force fields showed its capability to keep the positions of the binding sites in the selectivity filter stable after equilibration time.

The system has been initially fully equilibrated for 1.2 ns using GROMACS, in which the first 200 ps were useful to heat the structure from 100 to 300 K. After equilibration the simulation box is approximately 6.9 × 6.9 × 9.7 nm; then a further short equilibration (a few hundred picoseconds) was performed after the introduction of the biasing potential to obtain a starting configuration for each of the selected techniques and respective MD codes.

Electrostatic interactions have been computed using the smooth particle-mesh Ewald (PME) algorithm[18] implemented in the two codes with a fourth-order interpolation function, a grid of 72 × 72 × 100 points (corresponding to a mesh of less than 0.1 nm wide), and a cutoff in the direct and reciprocal space of 1.2 nm and 0.042 nm^{-1}, respectively. Considering the nanometric width of the selectivity filter, we also tested higher-order functions. However, a finer treatment of the PME did not add any further contribution or improvement to the present results and does not justify the increased computational burden.

All simulations have been carried out in the NVT ensemble, using the Nose-Hoover temperature coupling (reference temperature 300 K, time constant between 3 and 5 ps). A time step of 1 fs was used in GROMACS runs; the r-RESPA algorithm[19] was used in ORAC in conjunction with the rattle-shake algorithm to fix covalent bonds involving hydrogen atoms. The time steps used in the RESPA algorithm were 0.5-1-2-4 and 12 fs.

2.2. Computational Strategies To Bias MD Simulations. In the following a short overview of the three techniques applied in our calculations to bias the system toward the transitions of interest is presented, with the purpose of pointing out the theoretical assumptions and the computational strategies, which can possibly produce qualitative and quantitative effects on the results. Details are also provided about the values of the parameters used in our simulations for the KcsA case. These parameters have been fixed in order to optimize the convergence of the methods used to study the physical system at hand, but the sensitivity of each technique on its parameter set is beyond the scope of this work and it was not tackled with detail.

In the following we use PMF as a synonym for free-energy profile as a function of a set of chosen coordinates, which are also called reaction coordinates.

2.2.1. Umbrella Sampling. The problem of calculating the PMF is tackled with US technique[20] by adding a fictitious term to the Hamiltonian $H(\mathbf{x},\mathbf{p})$ of the system under investigation

$$\tilde{H}(\mathbf{x},\mathbf{p}) = H(\mathbf{x},\mathbf{p}) + h_i(r) \tag{1}$$

where **x** and **p** are the positions and the momenta of all the atoms of the system of interest.

This term is a static harmonic biasing potential, function of a chosen reaction coordinate $r = r(\mathbf{x})$, and center around at a given position r_i:

$$h_i(r) = \frac{1}{2} k(r - r_i)^2 \tag{2}$$

$h_i(r)$ is used to restrain the reaction coordinate r in the neighborhood of r_i, thus enhancing sampling of that region of the configuration space.

The center position r_i of the biasing potential is varied step by step along a defined path to obtain a set of N partially overlapping windows, each of them providing an ion probability distribution function $\rho_i(r)$. These distributions are then combined together to give the unbiased PMF by means of the weighted histogram analysis method (WHAM).[21]

Following the scheme suggested by Souaille and Roux,[22] the unbiased total distribution probability $\rho^u(r)$ is defined as

$$\rho^u(r) = C \sum_{i=1}^{N} \frac{n_i \exp[-\beta(h_i(r) - f_i)]}{\sum_{j=1}^{N} n_j \exp[-\beta(h_j(r) - f_j)]} \rho_i(r) \tag{3}$$

where $\beta = 1/k_BT$, C can play the role of a normalization constant, n_i is the weight of the simulation of the ith window, i.e., the number of configuration samples used to compute $\rho_i(r)$, and coefficients f_i, coming from the adding of the biasing potential, are calculated by an iterative solution of the formula

$$\exp(-\beta f_i) = C \int \mathrm{d}r \sum_{k=1}^{N} \frac{n_k \exp(-\beta h_i(r))}{\sum_{j=1}^{N} n_j \exp(-\beta(h_j(r) - f_k))} \rho_i(r) \tag{4}$$

The PMF is finally calculated by means of a generalization of the reversible work theorem[23]

$$G(r) = G(r_0) - k_BT \ln \frac{\rho^u(r)}{\rho_0^u(r_0)} \tag{5}$$

where r_0 is a reference point.

If more than one reaction coordinate is used in the simulation, the overall biasing potential is given by a sum of terms of the type reported in eq 2; in this case the center positions of the biasing potentials are varied on a grid on a multidimensional surface.

2.2.2. Steered Molecular Dynamics. SMD as well makes use of an added biasing potential to force the system to visit high free-energy regions. Contrary to the US technique, where the system is driven to the region identified by r_i (eq 2) and there statically remains until the following window is considered, in this case the added potential is continuously varied in time until the system reaches its ending configuration.

Therefore eq 1 is still valid, but in this case the new Hamiltonian is also a function of time that appears explicitly in the expression of the new biasing potential

$$h(\mathbf{x}, t) = \frac{1}{2} k[r(\mathbf{x}) - (\mu_0 + \upsilon t)]^2 \tag{6}$$

where μ_0 is the initial center position of the restraining potential, and v is the pulling velocity. This framework resembles atomic-force microscope experiments where a molecule is pulled between two positions, being subject to a time-varying external force.

The evaluation of the PMF relies on the Jarzynski's identity.[24] This equality links the equilibrium free-energy differences ΔG between the states A and B to the work W done on the system through all the nonequilibrium processes leading it from A to B. According to the second law of thermodynamics, ΔG represents the lower limit of $\langle W \rangle$: $\langle W \rangle \geq \Delta G$, the equality being valid in the limit of quasi-static (or equilibrium) processes. $\langle\rangle$ denotes an ensemble average. Jarzynski, however, proved that the following equality holds true regardless of the speed of the process:

$$\langle e^{-\beta W} \rangle = e^{-\beta \Delta G} \qquad (7)$$

The general validity of eq 7 depends on a small number of trajectories where $W_i \leq \Delta G$. The probability of these events decreases exponentially as the speed of the process increases, thus a large number of simulations is needed to handle a reliable statistical ensemble with even a relatively high pulling velocity. In practice, despite its theoretical speed-free validity, the applicability of this equation is limited to slow processes, whose energy fluctuations are comparable to k_BT, and a number of trajectories have to be combined together to obtain significant results.

By means of eq 7, one gets the free-energy of the system described by $\tilde{H}$ that must be corrected by subtracting the term due to the perturbing potential. Following the procedure described by Hummer and Szabo,[25] one finally gets the expression for the free-energy $G(r)$, as a function of the chosen reaction coordinate

$$G(r) = -\frac{1}{\beta} \ln \frac{\sum_t \frac{\langle \delta(r - r_t) \exp(-\beta w_t) \rangle}{\langle \exp(-\beta w_t) \rangle}}{\sum_t \frac{\exp[-\beta h(r,t)]}{\langle \exp(-\beta w_t) \rangle}} \qquad (8)$$

where the two summations are over time steps t, $h(r,t)$ is the perturbing potential defined in eq 6, and w_t is the work done on the system until time t. Note that for this kind of simulation a single reaction coordinate is used, so that a one-dimensional analysis is performed.

An improvement of SMD results can in principle be obtained by the Crooks equation[26] that makes use of both the forward and the backward average work and extends the Jarzynski identity. Whenever the hypotheses of the transient fluctuation theorem are satisfied, the works of the forward and backward transitions can be mixed together to give a better estimate of the PMF profile.

2.2.3. Metadynamics. MetaD[8,27−29] is a recently introduced technique based on the idea of the complexity reduction, being able to speed up the evolution of some defined reaction coordinates $r_k(\mathbf{x})$ with the introduction of a "history-dependent" biasing potential $V(r_k,t)$. The latter is the sum of repulsive functions that are added at given time steps during the simulation in order to constitute a "penalty" term for configurations already visited in the space of the reaction coordinates. These repulsive functions fill the minima of the PMF and, after a long simulation, tend to compensate exactly the underlying PMF that, in turn, can be approximated by their sum. If Gaussians are used as repulsive functions for the potential, then eq 1 rewrites

$$\tilde{H}(\mathbf{x}, \mathbf{p}, t) = H(\mathbf{x}, \mathbf{p}) + \sum_i w_i \exp\left[-\sum_{k=1}^{N} \frac{(r_k(\mathbf{x}, t) - r_k(\mathbf{x}_i, t_i))^2}{\Delta r_k}\right] \qquad (9)$$

where w_i and Δr_k are the height of the repulsive potential and the scale factor for the kth coordinate, respectively. The outer summation (index i) is over time steps. The scale factor defines the range of action of the repulsive potential and represents a sort of resolution of the reconstructed PMF.

The main advantage of MetaD with respect to US is that it is not required to define a priori the range of variation of the reaction coordinates, letting the system evolve toward the lowest transition state, thus obtaining the minimum free-energy landscape along the path connecting the two minima. This prevents the sampling of uninteresting regions, and, in principle, it allows the introduction of a high number of reaction coordinates. Similar approaches have previously been exploited to explore the configuration space, such as the taboo search,[30] and the local elevation method.[31] Moreover, MetaD can also be considered as an extension of the Wang-Landau algorithm,[32] and it is closely related to the recent adaptive force-bias algorithm,[33,34] where the derivative of the free-energy along a reaction coordinate is reconstructed by means of an adaptive time-dependent bias.

However, the use of a time-dependent biasing potential is in some way a nonequilibrium procedure with respect to the other degrees of freedom, especially to the so-called slow modes. When the latter are not included in the chosen set of reaction coordinates, the choice of the parameters controlling the repulsive potentials (i.e., deposition time step, height and scale factor) is crucial to let the system equilibrate each time a new term is added. The efficient sampling of nonexplicit slow modes within MetaD can be tackled in different ways, either improving the sampling with the replica exchange method,[35] or by means of the bias-exchange metadynamics,[36] or correcting the reconstructed PMF with a subsequent refining US.[37]

2.3. Choice of the Reaction Coordinates. A computational mapping of the free-energy profile for potassium ions in the KcsA protein was already done in the past[3,10] by means of multidimensional US. In that case, the authors used a number of occupancy configurations corresponding to the case of two ions within the selectivity filter, independently varying the position of these two ions together with that of a third ion in the cavity. More than 300 simulations were needed (leading to an aggregate total simulation time of 36 ns) which, in turn, implied a significant computation time. The obtained free-energy maps confirmed that the conduction process takes place as several consecutive steps in which, if the two ions within the filter move, they always move concertedly. This picture of the permeation process suggests the possibility of using a single curvilinear coordinate to describe ion motion within the US framework, thus avoiding to explore paths along which the system will never evolve due to excessively high energy. A further confirmation of what above stated can be deduced by analyzing the PMF plots reported in the reference work by Bernèche and Roux.[10] It can be observed that the preferred pathways can be split into segments where only one of the selected coordinates moves significantly, the others being confined in a narrow well centered on their initial value. The identification of a curvilinear coordinate to reduce the amount of CPU time is possible, though not trivial at all because it must not force unphysical movements of the ions.

The SMD technique seems to be suitable for a similar procedure. Under physiological conditions the ion flux through the channel is driven by a transmembrane potential, resulting from a charge imbalance between intra- and extracellular environments. As a consequence a potential that changes along the channel axis should exist, and the reaction coordinate should follow this observation. A natural choice of reaction coordinate is thus the z value of an ion's coordinate in a Cartesian orthogonal reference system, also taking into account the strong ion confinement in the xy plane.

Three simulations have been performed, using the position of the top outermost ion (initially labeled as K_2), the position of the middle ion (K_4), or the position of their center of mass (K_{com}) as reaction coordinates, respectively. We would like to point out again that this technique implicitly accounts for one-dimensional analysis, bounding the position of one ion (or group of ions) to the pulling spring and, in practice, forcing in this way the ions' motion. An analogous choice of reaction coordinates has also been performed for the US simulations, for comparison purposes.

The case of MetaD requires a different approach. The chance to sample *at the same time* different reaction coordinates together with the use of a history-dependent potential term automatically drives the dynamics of the system along the minimum energy path, thus avoiding the exploration of undesired regions, where one can suppose a priori that the system will hardly pass through. The reader easily understands that the choice of a suitable minimal set of coordinates is far from being a trivial point. In order not to waste time, coordinates must be independent from each other and represent a minimum set able to describe the evolution of the system under investigation, including the slow-modes. Preliminary investigations focused on the definition of the appropriate coordinates are often needed. In this case we found that the most effective set of coordinates is represented by the positions of the ion in the cavity K_{cav} and of the middle ion in selectivity filter K_4. It should also be noticed that a fair comparison of the US and/or SMD sampling with MetaD under equivalent conditions implies to project the n-D free-energy profile from MetaD along the minimum-energy path, i.e., as a function of a single curvilinear coordinate making use, for instance, of the nudged elastic band (NEB) method.[38] A second benefit of MetaD is represented by the possibility to introduce coordinates not linked to the physical position of the ions, e.g., the water coordination number of the ion in the cavity. It is likely to suppose that the hydration/dehydration process affecting the ion in the cavity may be significant in the permeation process as much as the position of the ions in the selectivity filter. This concept has been underlined in different channels by Gervasio et al.[39] and Braun-Sand et al.[40]

2.4. Parameters and Computational Details. One of the major points concerning the comparison among different techniques aiming at the same result is the proper choice of simulation parameters for each technique to avoid that one technique outperforms the others only because of an unfair set of parameters.

The state of the art of US simulations for the KcsA channel is reported by Bernèche and Roux,[10] where the system was investigated with great detail, and it is used as a guide in the following. We used a force constant $k = 8368$ kJ/mol nm^2 (20 kcal/mol Å^{-2}) and a step between two consecutive centers of the biasing potential (r_i) $\Delta r = 0.05$ nm to respect the overlapping constraint. Up to 18 steps, 515 ps-long each, have been combined together, and the biased distributions have been reconstructed by 250 bins ranging from $3{\cdot}10^{-3}$ to $3.8{\cdot}10^{-3}$ nm in width, depending on the chosen reaction coordinate, as described in the next subsection.

The first 15 ps of each simulation were used to adjust the biasing potential to the new position, starting from the previous configuration. Then the following 250 ps were discarded as equilibration time in the presence of the biasing potential. The remaining 250 ps were then split into 5 blocks of 50 ps, and, finally, the results were averaged. The reaction coordinate was recorded at every time step, for a total productive sampling roughly about 2 million configurations per reaction coordinate. By this protocol it was possible to calculate the statistical error of the estimated PMF, which results to be about $\pm 1\ k_{\mathrm{B}}T$ at room temperature (corresponding to 2.5 kJ/mol).

In the absence of previously published studies on the KcsA channel with SMD, we performed several preliminary tests to determine both a suitable force constant and a pulling velocity. In a work of Jensen et al. on the permeation of glycerol through aquaglyceroporin GlpF[41] the SMD technique was intensively used to compute the energetics, and the influence of its parameters on the final result was also revised with details.

Following that suggested scheme, a harmonic constraint with a spring constant $k = 1673.6$ kJ/mol nm^2 (4 kcal/mol Å^{-2}) was attached to the selected reaction coordinate. This latter constant ensures a thermal fluctuation of the constrained coordinate of about $\sqrt{k_{\mathrm{B}}T/k} = 0.04$ nm and a corresponding force fluctuation of approximately 100 pN. We determined that a pulling velocity $\upsilon = 1{\cdot}10^{-3}$ nm/ps is slow enough to guarantee that the system always evolves through intermediate quasi-equilibrium states. This was further confirmed by performing reverse transitions at a double steering velocity and by observing that they converge to the same value of the free-energy barrier (see also Figure 3).

A preliminary run keeping the perturbing potential fixed was performed. The first 150 ps were discarded, then system configurations were saved every 50 ps, as input starting configuration for subsequent productive runs. The 50 ps interval ensures that saved configurations are uncorrelated. Each productive run lasted 500 ps, and the reaction coordi-

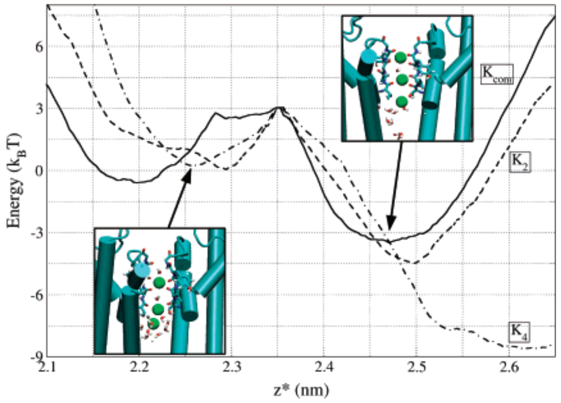


Figure 2. PMFs from US for the transition (S_2, S_4, S_{cav}) ↔ (S_1, S_3, S_{cav}). Solid, dashed, and dash-dotted lines refer to different reaction coordinates (see framed labels); abscissas have been shifted in order to allow direct comparison of the results. The maximum has been used as the pivotal point; the zero-level of the free energy is at arbitrary position. The two insets represent the configurations corresponding to the two minima.

nate was saved at each time step, for a total simulation time of 4.5 ns and 4 million sampled points. The PMF profile was reconstructed by means of eq 8 adapting the weighted histogram method[21] with 250 bins combining eight uncorrelated trajectories together. That proved to be enough for convergence. The statistical uncertainty on the PMF was also investigated by averaging 7 blocks of 8 trajectories, thus obtaining an error of ±2 k_BT (5 kJ/mol) on the energy scale and ±0.05 nm for the position of the minimum.

MetaD simulations were performed by means of two reaction coordinates, namely the positions of two ions in the filter-cavity region along the channel axis. We adopted the following protocol for the simulations: 2 kJ/mol-high hills were added every 4 ps, and the scale factor of reaction coordinates was set in order not to exceed the thermal fluctuations of the two coordinates in absence of any bias. Values of 0.02 and 0.015 nm looked appropriate for the ion in the cavity and the ions inside the filter, respectively. The latter value is less than the former due to their reduced mobility, as also stated in ref 14. A previous metadynamics study on the motion of ions inside chloride channels[13] guided this choice and actually ensures that the error on the reconstructed PMF is of the order of 2 k_BT (5 kJ/mol). The overall MetaD simulation lasted approximately 9 ns.

3. Results and Discussion

MD free-energy results usually depend both on the adopted biasing methodology and on the particular force field used in the simulations. An exhaustive comparison of different force-field models applied to gramicidin A as a test case can be found in the recent literature.[2]

The focus of our calculations is on the comparison among the three different biasing techniques illustrated in section 2 with the use of the GROMOS87 force field. It was recently proved that GROMOS*nn* force fields introduce a systematic overestimation of the energy barriers,[2] due to the peculiar set of electric charges included in their parametrization of the electrostatic interaction. This evidence, however, does not affect the validity of our conclusions because it equally influences all of the results under comparison. The obtained energy profiles fairly agree with those reported in the literature.[10]

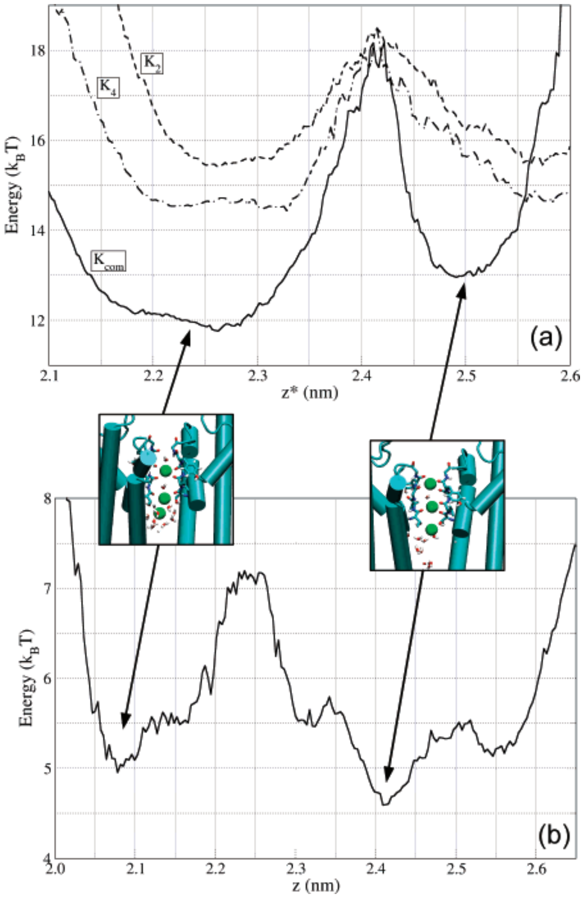


Figure 3. (a) (top) PMFs from SMD for the transition (S_2, S_4, S_{cav}) → (S_1, S_3, S_{cav}). Solid, dashed, and dash-dotted lines refer to different reaction coordinates (see framed labels); abscissas have been shifted in order to allow direct comparison of the results. The maximum has been used as the pivotal point; the zero-level of the free energy is at arbitrary position. The two insets represent the filter configurations corresponding to the two minima. (b) (bottom) PMF for the reverse transition (see text).

Furthermore, we point out that energy values are given using k_BT units. This choice is convenient since the results are used within the framework of the reaction-rate theory, with the purpose of studying conduction properties.[5]

3.1. US vs SMD. PMFs from US and SMD for transition (S_2, S_4, S_{cav}) → (S_1, S_3, S_{cav}) are reported in Figures 2 and 3a, respectively.

For comparison purposes, the curves within each figure corresponding to different reaction coordinates have been shifted using the free-energy relative maximum as the pivotal point for both *x*- and *y*-axes. Spatial differences in the position of minima and maxima obtained with US and SMD are limited within 0.1 nm or less, and they are attributed both to rigid shifts of the filter structure with respect to the internal reference during the simulation, and to thermal fluctuations. They do not affect either the PMF determination or the comparison between the two methods. Data are shown in the range 2.1−2.65 nm for US and 2.1−2.6 nm for SMD to include the abscissas corresponding to the initial and final configurations. The free-energy zero-level is at an arbitrary position in the two cases presented, which again does not influence the evolution of the energy difference separating the configurations.

Two minima are found, as expected, since the two corresponding configurations (S_2, S_4, S_{cav}) and (S_1, S_3, S_{cav}) of Figure 1 have been already identified as stable configurations of the selectivity filter.[10,11,42] All of the curves obtained are qualitatively similar, even though it should be remarked that the minimum of the final state in the US calculation is deeper than the one found with SMD. In this latter case, the second minimum is approximately as deep as the starting point.

In both plots two of the three curves (K_{com} and K_2 in US, K_2 and K_4 in SMD) are quite similar in shape and in numerical values, with differences less than 1 k_BT that can be easily associated with the statistical variance; nonetheless, the third curves (K_4 in US and K_{com} in SMD), though confirming a similar overall shape, present significantly different numerical values.

These differences are not related to computational uncertainties but instead to spatial fluctuations attributed to variations of the ion position within the cavity and to torsions of the residues facing the selectivity filter, as already observed in the past.[12] These changes, which take place on the time scale of our MD simulations, give origin to variations of the energy barriers larger than the computational uncertainty of MD but still small with respect to barriers separating conducting from nonconducting states. For this reason we have considered the previous results as a more likely estimate of the barrier height provided by the two techniques considered in this paragraph.

The interpretation of the results coming from the application of the SMD technique requires some care. With reference to Figure 3a, the method provides a reliable estimate of the energy barrier associated with the transition out of the minimum of the PMF, which is the starting point of the pulled atoms. After barrier crossing the initial spring length is not fully recovered, i.e., some elongation is still present. This fact produces an artificial overestimate of the energy level associated with the final state of the pulled transition. To get a correct estimate, the reverse transition, namely (S_1, S_3, S_{cav}) → (S_2, S_4, S_{cav}), has been simulated starting from an initially fully relaxed spring and pulling the ion K_3 toward site S_4. The results of this simulation, reported in Figure 3b, compare with the curve labeled K_4 in Figure 3a. Here the overestimate due to the pulling action of the spring does not allow a correct sampling of the minimum located at 2.07 nm. By analyzing the two results together we can conclude that, within the numerical uncertainty associated with the SMD method, the energy barriers associated with forward and reverse transitions result in being equal, in contrast to what is observed with US. From the US runs we obtain an estimate of the barrier of about 3 k_BT for the forward transition and of about 6 k_BT for the reverse transition (except curve K_4, where a 9 k_BT barrier is found), which means a deeper second minimum.

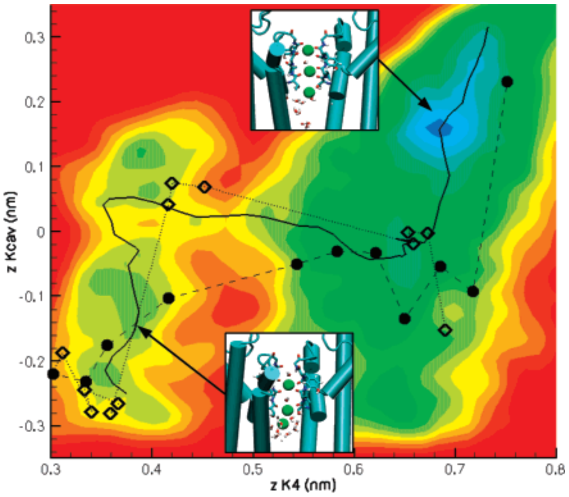


Figure 4. PMF from MtD for the transition (S_2, S_4, S_{cav}) ↔ (S_1, S_3, S_{cav}); each color level corresponds to an energy of 1 k_BT. The full line represents the minimum-energy path within MetaD; the dots and the diamonds are snapshots taken from trajectories followed by the curvilinear coordinate in US and SMD runs, respectively. The dotted and dashed lines are drawn to guide the eye and not as real paths. The two insets illustrate the position of the ions in the selectivity filter corresponding to the two main minima.

The comparison of the above results with those obtained with MetaD helps to justify this discrepancy.

Some enhancements may occur, and a more precise PMF profile can be achieved by means of the Crooks equation. However for the present case the error affecting SMD calculations is fair enough to let us infer that the strong qualitative difference existing among these profiles and those obtained by US and MetaD can only be attributed to the different number of sampled coordinates. The main disadvantage of SMD over US and MetaD is thus represented by the intrinsic one-dimensional behavior of eq 7 that leads to a unique coordinate analysis. When a multiplicity of coordinates is required, as the present case looks like, the SMD picture is clearly too poor.

3.2. MetaD vs US. MetaD results for the energy landscape involved in the transition (S_2, S_4, S_{cav}) → (S_1, S_3, S_{cav}) are reported in Figure 4 as functions of the position of ion K_4 and of ion K_{cav}. A 3 k_BT barrier is estimated by all of the selected techniques, and the general trend of a deeper minimum for the (S_1, S_3, S_{cav}) configuration, as obtained from US, is also confirmed. It is worth noticing that the two minima do not correspond to the same position of the ion in the cavity: when the latter moves closer to the near vacant site inside the selectivity filter the free-energy surface shows a deeper minimum.

To better link the results obtained with the three different methods we can analyze the trajectories followed by the moving ions in US and SMD runs and the time needed to observe the transition. Values reported in Figure 4 (and in Figure 6) must be interpreted as the time-averaged position of the K_4 and K_{cav} ions within US windows (K_4 trajectory) and as snapshots of a representative run in the SMD case. The standard deviation of ion positions range from 0.013 to 0.021 nm for K_4 and from 0.04 to 0.13 nm for K_{cav} as US is

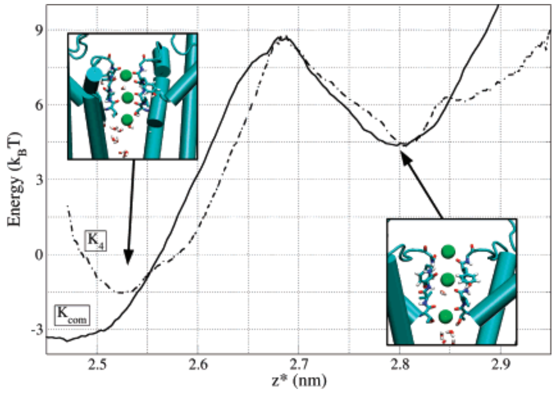


Figure 5. PMFs from US for the transition (S_1, S_3, S_{cav}) ↔ (S_0, S_2, S_4). Solid and dash-dotted lines refer to different reaction coordinates (see framed labels); abscissas have been shifted in order to allow direct comparison of the results. The maximum has been used as the pivotal point; the zero-level of the free energy is at arbitrary position. The two insets represent the configurations corresponding to the two minima.

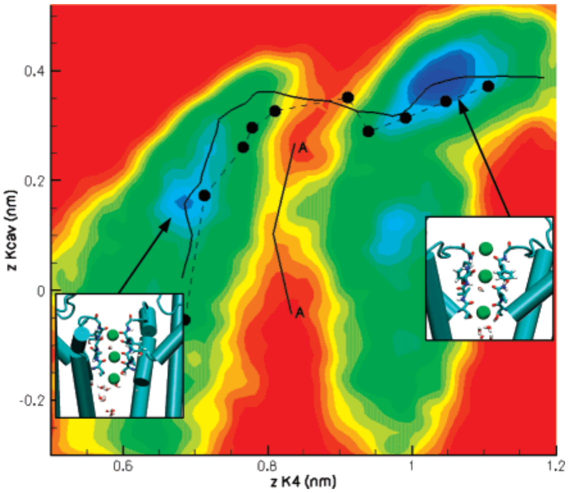


Figure 6. PMF from MtD for the transition (S_1, S_3, S_{cav}) ↔ (S_0, S_2, S_4); each color level corresponds to an energy of 1 k_BT. The full line represents the minimum-energy path within MetaD; the dots are snapshots taken from trajectories followed by the curvilinear coordinate in US (the dashed lines are drawn only to guide the eyes and do not correspond to a physical path). Line A−A represents a wall separating the two regions, the saddle point in that position is due to an unavoidable artifact of the interpolation surface. The two insets illustrate the configurations corresponding to the two minima.

concerned and approximately 0.05 nm (K_4) and 0.1 nm (K_{cav}) in the case of SMD. At the beginning, the ions reside in their original positions, which correspond to the first minimum in the energy plot; then, as we start pulling one of the ions in the selectivity filter with SMD, their energy is increased, and finally the barrier is crossed. Meanwhile, the ion in the cavity cannot rise up to the neighboring site in the selectivity filter until the transition takes place, and it is observed to strongly oscillate within the cavity. The configurations with the ion in the cavity and the ion close to site S_4 are in fact almost isoenergetic (Figure 4). After the transition, some time is needed for the ion in the cavity to stabilize and enter the selectivity filter; this, in turn, lets the system evolve toward its new free-energy minimum. The full translocation sequence can, in principle, be observed with US, but not with SMD, since the continuous pulling action in practice does not allow the system to reach the final equilibrium. A similar explanation can also be given for the deeper minimum of curve K_4 of US analysis (Figure 2): both curves labeled K_{com} and K_2 correspond to simulations where the ion in the cavity is not substantially moving. According to this interpretation, the role of ion K_4 is crucial for the conduction process, because it influences directly the position of the other ions in the cavity and in the selectivity filter.

To further confirm the hypothesis above, we have also investigated the transition (S_1, S_3, S_{cav}) → (S_0, S_2, S_4) with US and MetaD. SMD was not used any longer because it proved not to be able to correctly identify the position of the minimum corresponding to the (S_1, S_3, S_{cav}) configuration in the previous analyzed transition. For the situation at hand, it must be remarked that, if the selectivity filter is not populated by two ions, the protein undergoes a significant conformational change leading to a nonconductive state.[9] For this reason we always have to consider reaction coordinates directly or indirectly involving the intermediate ion in the filter. If we used the outermost ion position as the only reaction coordinate, we would have run the risk of emptying the selectivity filter and, thus, driving the system to a nonconductive state. On the other hand, the ion in the cavity can only fill site S_4: in the final configuration of the previous analyzed transition this ion resided quite close to the filter's mouth. Figures 5 and 6 show the results for US and MetaD, respectively. A general agreement from both a qualitative and a quantitative point of view is found between the two techniques: the path followed with 1-D US coordinate is the minimum energy path also identified within the MetaD framework, and the exit barriers do not differ very much from each other (10−12 k_BT for US, approximately 13 k_BT for MetaD). The MetaD analysis reveals two minima corresponding to transition (S_3 → S_2), depending on the final position of the ion originally in the cavity, which can either reside close to or fully enter site S_4. The most stable one is represented by the latter case, since it corresponds to a deeper minimum. The minimum-energy path calculated with the NEB method and reported in Figure 6 shows that the transition happens when the ion in the cavity enters site S_4 and not when it is adjacent to it, even if this corresponds to a slightly higher barrier. This interpretation is further confirmed by the fact that in the transition (S_1, S_3, S_{cav}) → (S_0, S_2, S_4) potassium ions can occupy two adjacent sites for short periods of time, breaking the rule of concerted motion. In particular this sequence of events, reported in Figure 7, takes place: the ion originally located in S_{cav} enters the site S_4, being S_3 filled; then, the water molecule in S_2 exchanges its position with the ion in S_3, leading to (S_1, S_2, S_4) configuration. The latter configuration is quite unstable (top of the barrier) and evolves to (S_0, S_2, S_4), which represents the final state of this transition. These intermediate three-ion states, in which potassium ions occupy two adjacent binding sites, have already been reported in the literature,[3] and they have recently been observed also in the analysis of the permeation paths of homologous channel Kv1.2.[43]

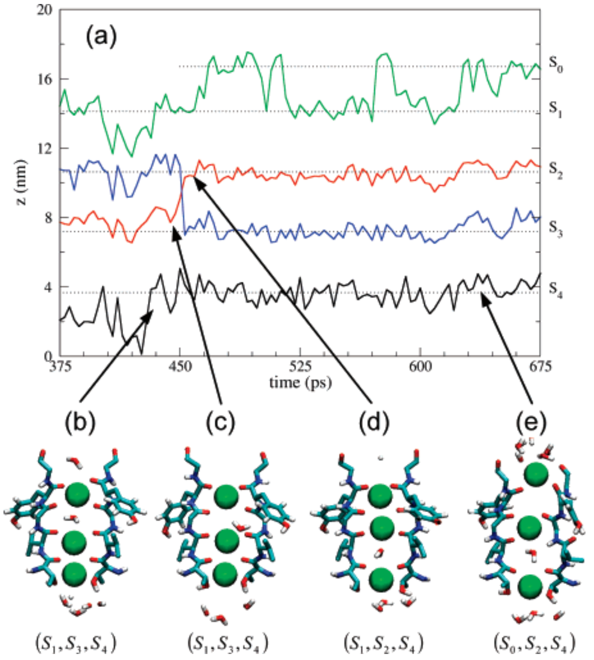


Figure 7. Trajectories of the potassium ions (green, red, and black lines) and of the water molecule (blue line) in the selectivity filter, projected onto the symmetry axis of the channel, during transition (S_1, S_3, S_{cav}) → (S_0, S_2, S_4) (a) and significant MD snapshots of this transition (b)−(e). For clarity purposes, only two subunits are represented. The positions of the binding sites are represented by dotted lines. Data are recorded every 3 ps. At around 450 ps the water molecule and the ion initially in S_3 exchange their position within the imposed time step: snapshots (c) and (d) suggest that the water molecule first moves off-axis close to the backbone, facilitating the ion to rise up into site S_2, and then slips down into its final position. A fast flip of a carbonyl oxygen of Val76, separating sites S_2 and S_3, is also possible on a shorter time scale. A similar reorientation is reported also in snapshot (e), when a hydrogen bond between the carbonyl oxygen of Val76 and the protein backbone behind it (not represented in the figure) is established.

Furthermore, we did not observe a substantial rearrangement of the selectivity filter in connection with the water−potassium exchange between two subsequent simulation snapshots (10 ps); however, we cannot exclude that this may take place on a shorter time scale. The overall PMF calculated via MetaD and projected onto the minimum-energy path is reported in Figure 8.

When the (S_0, S_2, S_4) configuration is reached, we have observed a symmetry breaking, also reported in a previous work by Bernèche and Roux.[44] As it is shown in Figure 7e, the amide plane Val76-Gly77 undergoes a 180° reorientation, while the backbone carbonyl oxygen of Val76 points away from the conduction path. As a consequence the water molecule in site S_3 forms a hydrogen bond with the hydrogen of Gly77. At the same time the plane of the aromatic ring of Tyr78 bends down, getting closer to the reoriented carbonyl oxygen of Val76, but it keeps its own specific mobility preserved.

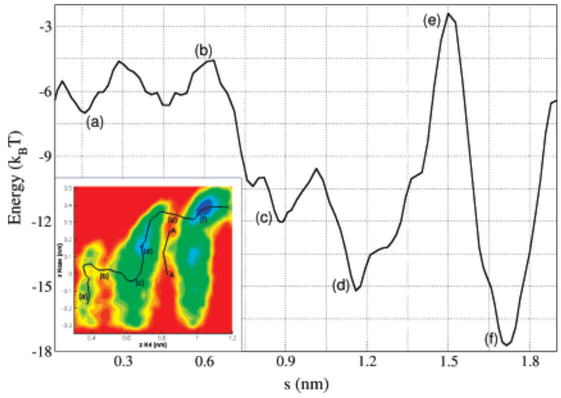


Figure 8. MetaD PMF projected onto the minimum-energy path by means of the NEB method. Data are presented as a function of a curvilinear coordinate, and labels are used to identify corresponding points in the free-energy landscape. In the inset the whole PMF landscape and trajectory are also reported.

In conclusion, both MetaD and US allow the easy use of many reaction coordinates, even though with different sampling strategies: MetaD performs a simultaneous sampling of the whole set of reaction coordinates, while within the US framework only one coordinate is varied at a time, the others being temporarily kept fixed as parameters. As a consequence, MetaD is less resource-demanding than US and provides results faster. In our case the 2D plot obtained with MetaD in Figures 4 and 6 would have required up to ten times the simulation time used, if calculated using US.

US, however, provides a more accurate sampling than MetaD when the same set of coordinates is considered, in particular when the PMF exhibits many competing pathways in the explored area. In our case, for example, Figure 6 can be compared with Figure 2 on the left in the reference work by Bernèche and Roux.[10] A general qualitative (and, to a less extent, even quantitative) agreement between the two PMFs is found, but some differences exist. The free-energy path explored by MetaD corresponds to the major path found by US, but the secondary path identified by US has not been mapped by MetaD. This secondary path can be associated with either a slightly higher energy or slow modes not sampled by the chosen set of reaction coordinates. It should be noticed, however, that the region corresponding to the (S_0, S_2, S_{cav}) configuration is visited after that (S_0, S_2, S_4) is reached, thus warning about the possible existence of a competing path. The saddle point linking (S_1, S_3, S_{cav}) to (S_0, S_2, S_{cav}) is an artifact of the interpolating surface, as the NEB analysis testifies. Last, it must also be pointed out that the (S_1, S_3, S_4) configuration corresponds to an intermediate state along the transition path close to the energy maximum. The associated secondary minimum reported in the reference work by Bernèche and Roux is not revealed by the MetaD simulation.

Both techniques seem to be successful and accurate enough (within their own constraints and limitations) to describe the permeation process of an ion in a narrow pore. A statement about which of the two outperforms the other strictly depends also on the system under investigation, and, for this reason, the choice of MetaD instead of US (or vice versa) must be done after a careful evaluation of the trade-off level between computational efficiency and sampling accuracy.

4. Conclusion

A compared analysis of three different numerical techniques (Umbrella Sampling, Steered Molecular Dynamics, and Metadynamics) aimed at the reconstruction of the free-energy landscape via molecular dynamics has been presented. As a case study, two key transitions of the permeation process in the KcsA channel have been chosen.

The obtained results suggest few conclusive statements, which can be considered general and applicable also to other permeation cases in nanometric pores.

All of the three techniques represent computational tools able to reconstruct the PMF profile composed by a number of valleys and hills of different height. From a qualitative point of view the identification of the minimum energy path connecting two consecutive valleys, i.e., finding the lower barrier existing between two ion occupancy configurations, is a straightforward activity that can be performed with an average error of few k_BT. Barriers giving origin to an effective permeation path must, in fact, be smaller than the ones existing among configurations not involved in the conduction. However, the statistical errors allow only approximate quantitative estimate of transport properties, such as, for instance, the ionic current. One more critical point is represented by the choice of the modeling force field that can introduce a further systematic source of uncertainty.

With regards to the three selected techniques, SMD proved to be the less suitable one. In fact, this technique was originally developed for studying the stretching and/or unfolding of proteins and, then, adapted to investigate free-energy landscapes. This technique provided poor results for the present case: when the PMF cannot be described by means of a unique physical reaction coordinate, SMD may be inaccurate because the continuous pulling of the reaction coordinate prevents the other coordinates from reaching values leading to the global energy minimum.

The US technique is the straightforward, most used, and most accurate technique for this kind of analyses. However it demands intensive computational efforts when many coordinates are used. MetaD is faster in achieving results, once an appropriate set of reaction coordinates is chosen. Moreover, it also allows the use of nonintuitive reaction coordinates not directly related to the spatial coordinates of the permeating ions (e.g., ion hydration). For this reason it can be preferred as computationally advantageous. With respect to US, MetaD tends to drive the simulation through the main paths of the free-energy profile, whereas US provides a description of the whole landscapes, investigating also secondary pathways that could be accessed via MetaD only by increasing the number of variables and, consequently, the computational burden. Furthermore, MetaD provides an upper limit to barriers, because it may happen that some slow modes are not explicitly taken into account. On the other hand, US provides a lower limit, due to the fact that the transition state itself is always considered an equilibrium distribution by the computational procedure. The combined use of MetaD and US, the former being able to scope out the dominant reaction coordinates and the latter to refine results, represents a good suggestion to achieve accurate results with an affordable computational cost.

Acknowledgment. We gratefully acknowledge the CINECA National Supercomputing Center for providing CPU time and the CASPUR Computer Center (Rome) for additional CPU time allocation. The Italian Ministry for University and Research (MIUR) supported funding through project PON-Cybersar.

References

(1) Roux, B.; Allen, T. W.; Bernèche, S.; Im, W. *Quart. Rev. Biophys.* **2004**, *37*, 15.

(2) Allen, T. W.; Andersen, O. S.; Roux, B. *Biophys. J.* **2006**, *90*, 3447.

(3) Bernèche, S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8644.

(4) Mapes, E. J.; Schumaker, M. F. *Bull. Math. Biol.* **2006**, *68*, 1429.

(5) Piccinini, E.; Affinito, F.; Brunetti, R.; Jacoboni, C.; Rudan, M. *J. Chem. Theory Comput*. **2007**, *3*, 248.

(6) Gullingsrud, J. R.; Braun, R.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 190.

(7) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275.

(8) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562.

(9) Zhou, Y.; Morais-Cabral, J.; Kaufman, A.; MacKinnon, R. *Nature* **2001**, *414*, 43.

(10) Bernèche, S.; Roux, B. *Nature* **2001**, *414*, 73.

(11) Aqvist, J.; Luzhkov, V. *Nature* **2002**, *404*, 881.

(12) Bernèche, S.; Roux, B. *Biophys. J.* **2002**, *78*, 2900.

(13) Gervasio, F. L.; Parrinello, M.; Ceccarelli, M.; Klein, M. L. *J. Mol. Biol.* **2006**, *361*, 390.

(14) Compoint, M.; Carloni, P.; Ramseyer, C.; Girardet, C. *Biochim. Biophys. Acta* **2004**, *1661*, 26.

(15) Lindhal, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.

(16) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.

(17) Procacci, P.; Paci, E.; Darden, T. A.; Marchi, M. *J. Comput. Chem.* **1997**, *18*, 1848.

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

(19) Tuckerman, M. E.; Berne, B. J.; Matryna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990.

(20) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.

(21) Kumar, S.; Bouzida, D.; Swensen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(22) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40.

(23) Chandler D. Statistical Fluids. In *Introduction to modern statistical mechanics*; Oxford University Press: Oxford, New York, 1987; pp 188−233.

(24) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(25) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3658.

(26) Crooks, G. E. *J. Stat. Phys.* **1998**, *90*, 1481.

(27) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302-1.

(28) Ceccarelli, M.; Danelon, C.; Laio, A.; Parrinello, M. *Biophys. J.* **2004**, *87*, 58.

(29) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714.

(30) Cvijovic, D.; Klinowski, J. *Science* **1995**, *267*, 664.

(31) Huber, T.; Horda, A.E.; van Gunsteren, W. F. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695.

(32) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(33) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *108*, 1964.

(34) Henin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904.

(35) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435.

(36) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553.

(37) Babin, V.; Roland, C.; Darden, T. A.; Sagui, C. *J. Chem. Phys.* **2006**, *125*, 204909.

(38) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; pp 385−403.

(39) Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2005**, *127*, 2600.

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

(41) Jensen, M. Ø.; Park, S.; Tajkhorshid, E.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6731.

(42) Morais-Cabral, J. H.; Zhou, Y.; MacKinnon, R. *Nature* **2001**, *414*, 37.

(43) Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. *Biophys. J.* **2006**, *91*, L72.

(44) Bernèche, S.; Roux, B. *Structure* **2005**, *13*, 591.

CT7001896

JCTC Journal of Chemical Theory and Computation

Approach for the Simulation and Modeling of Flexible Rings: Application to the α-D-Arabinofuranoside Ring, a Key Constituent of Polysaccharides from *Mycobacterium tuberculosis*

Mikyung Seo,† Norberto Castillo,† Robert Ganzynkowicz,† Charlisa R. Daniels,‡
Robert J. Woods,‡ Todd L. Lowary,*,† and Pierre-Nicholas Roy*,†

Department of Chemistry and Alberta Ingenuity Centre for Carbohydrate Science,
Gunning-Lemieux Chemistry Centre, University of Alberta,
Edmonton AB T6G 2G2, Canada, and Complex Carbohydrate Research Center,
University of Georgia, 315 Riverbend Road, Athens, Georgia 30602

Received October 25, 2007

Abstract: A number of lower organisms (bacteria, fungi, and parasites) produce glycoconjugates that contain furanose rings. Of particular interest to our group are cell wall polysaccharides from mycobacteria, including the human pathogen, *Mycobacterium tuberculosis*, which contain a large number of arabinofuranose resides. As part of a larger project on the conformational analysis of these molecules, we report here molecular dynamics simulations on methyl α-D-arabinofuranoside (**1**) using the AMBER force field and the GLYCAM carbohydrate parameter set. We initially studied the ability of this method to predict rotamer populations about the hydroxymethyl group (C4−C5) bond. Importantly, we show that simulation times of up to 200 ns are required in order to obtain convergence of the rotamer populations for this ring system. We also propose a new charge derivation approach that accounts for the flexibility of the furanoside ring by taking an average of the charges from a large number of conformers across the psuedorotational itinerary. The approach yields rotamer populations that are in good agreement with available NMR data and, in addition, provides insight into the nature of the puckering angle and amplitude in **1**.

Introduction

Furanose rings are important components of a number of glycoconjugates, with the most well-known examples being the nucleic acids, which contain either D-ribofuranose or 2-deoxy-D-*erthyro*-pentofuranose (2-deoxy-D-ribose).[1] It is less widely appreciated that a number of bacteria, fungi, and parasites also biosynthesize furanoside-containing polysaccharides.[2,3] Glycans composed of furanosyl moieties are typically found on the surfaces of the organisms that produce them, and thus they play important roles in the interaction of these species with their environment. Furanosyl residues are also key components of natural products other than polysaccharides, including plant opines,[4] glycopeptides,[5] and the aminoglycoside antibiotics.[6]

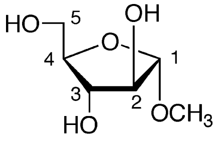
Among the most elaborate examples of these glycoconjugates are two polysaccharides, arabinogalactan (AG) and lipoarabinomannan (LAM), that are found in the cell wall of mycobacteria.[7] Notable among these are the pathogenic organisms *Mycobacterium tuberculosis*, *M. leprae*, and *M. avium*, which, respectively, cause tuberculosis, leprosy, and a tuberculosis-like disease common in HIV-positive individuals. The AG, a polysaccharide containing approximately 100 monosaccharide units, is composed entirely of arabinofuranose and galactofuranose residues, except for two pyranose moieties, which serve as the linker between the

* Corresponding author e-mail: pn.roy@ualberta.ca (P.-N.R.), tlowary@ualberta.ca (T.L.L.).

† University of Alberta.

‡ University of Georgia.

10.1021/ct700284r CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/14/2007



1

Figure 1. Structure of methyl α-D-arabinofuranoside.

glycan and peptidoglycan.[8] Similarly, a significant component of LAM is an arabinan domain, representing approximately half the molecular weight, which contains only arabinofuranose residues.[9]

The AG is esterified at its nonreducing end with mycolic acids, $C_{70}-C_{90}$ branched lipids, yielding a glycolipid named the mycolyl-arabinogalactan (mAG) complex, which is the major structural component of the cell wall. In the accepted model for the macrostructure of the mycobacterial cell wall,[10] the mycolic acids pack perpendicular to the plasma membrane thus forming a lipid layer at the periphery of the assembly. This layer of tightly packed mycolic acids serves as a major permeability barrier to the passage antibiotics. Thus, the AG serves as a scaffold by which the organism attaches an additional permeability barrier to peptidoglycan and, in turn, the plasma membrane.

The essentially exclusive presence of furanosyl rings in the AG is curious as they are of higher energy than their pyranose counterparts.[11] It has been hypothesized[12] that the AG is composed of furanose residues because the resulting glycan has greater flexibility relative to a pyranose-containing species. A more malleable scaffold would be expected to facilitate optimal packing of the mycolic acids, which in turn would provide the organism with great protection against its environment. This "flexible-scaffold hypothesis" is plausible considering that while pyranose rings typically adopt single well-defined chair conformations, furanose rings can exist in a variety of twist and envelope conformations that are separated by typically low-energy barriers. Therefore, a polysaccharide composed of furanose residues would be more flexible than one made of pyranose residues. Although intriguing, there are scant data to support the flexible scaffold hypothesis. As part of a program dedicated to understanding the conformation of mycobacterial AG (and LAM), we have carried out a series of NMR studies on the arabinofuranose-containing oligosaccharides[13,14] and coupled these experimental studies with high-level ab initio and density functional theory calculations on methyl α-D-arabinofuranoside (**1**, Figure 1)[15,16] and related analogs.[17−20]

Given their inherent flexibility, the conformational analysis of furanosides is more complicated than comparable studies with pyranosides as more than one ring conformer must be considered. For a given furanoside, the ten unique envelope (E) and twist (T) forms can be identified on the pseudorotational wheel (Figure 2), each with a unique psuedorotational phase angle, *P*.[21] In solution, there is a dynamic equilibrium between ring conformers, usually dominated by two major species between which the interconversion barrier is low (typically <5 kcal/mol).[17] One of these conformers is generally found in the northern hemisphere of the pseudorotational wheel, and the other in the southern hemisphere,[21] which are termed the North and South conformers, respectively. Conformational investigations of furanoside rings by NMR spectroscopy most commonly involve analysis using PSEUROT,[22] a program that assumes this two-state equilibrium and which fits the experimental $^{1}H-^{1}H$ coupling constant data to two conformers and their populations.

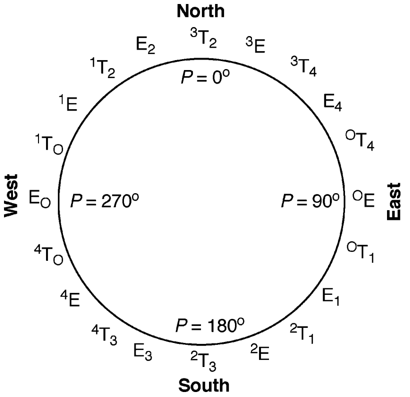


Figure 2. Pseudorotational itinerary for a D-aldofuranose ring.

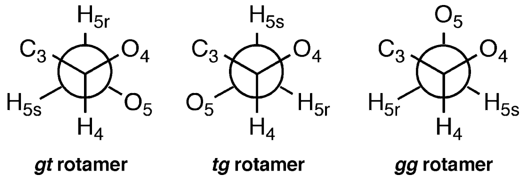


Figure 3. Definition of *gg*, *gt*, and *tg* rotamers about the C4−C5 bond.

Other key conformational features of importance include rotamer populations about the glycosidic C1−O1 and C4−C5 bonds. The preferred rotamer about the C1−O1 bond is the one in which the aglycone (e.g., the methyl group in **1**) is oriented *anti* with respect to the C1−C2 bond, as this is favored by the *exo*-anomeric effect.[23] For the C4−C5 bond (ω angle), three rotamers are typically present, *gt*, *tg*, and *gg* (Figure 3), with the distribution being influenced by a combination of steric and stereoelectronic (*gauche*) effects.[24−27]

Having studied the conformation of **1** using both experimental and high-level computational methods, we are interested in looking at larger oligomers of D-arabinofuranose, for which we have NMR data.[13,14] However, given the size of these molecules, their treatment with ab initio or density functional theory methods is of limited practicality. Thus, we have begun to investigate the use of force field models to probe the conformation of these oligosaccharides. Previous molecular mechanics studies of furanosyl rings have largely been carried out using MM3 or earlier variants of this force field.[28−34] However, over the past several years the use of the AMBER[35] force field in conjunction with the GLYCAM[36,37] parameter set has emerged as a reliable force field for molecular mechanics studies of oligosaccharides containing pyranose rings. In this paper, we describe the results of our first investigations of the use of the GLYCAM parameters and the AMBER force field to study the conformation of furanoside rings. More specifically, we report the ability of this computational method to predict the rotamer distribution about the C4−C5 bond and pseudorotational phase angle in **1** as determined by NMR spectroscopy. In this regard,

Table 1. Partial Atomic Charges of **1** Obtained Using the Usual GLYCAM Procedure for Five Reference Rings (A−E) and Using the Averaged Approach Described Here[a,b]

atom	A	B	C	D	E	ring averaged[c]
$P = 13$	$P = 13$	$P = 32$	$P = 139$	$P = 58$	$P^* = 31$	
$\phi_m = 34$	$\phi_m = 41$	$\phi_m = 40$	$\phi_m = 35$	$\phi_m = 40$	$\phi_m^* = 35$	
C1	0.38 (0.05)	0.37 (0.06)	0.38 (0.05)	0.37 (0.04)	0.38 (0.05)	0.38 (0.04)
C2	0.35 (0.09)	0.33 (0.09)	0.30 (0.07)	0.31 (0.05)	0.28 (0.09)	0.31 (0.07)
O2	−0.72 (0.02)	−0.73 (0.02)	−0.69 (0.02)	−0.70 (0.02)	−0.70 (0.03)	−0.69 (0.02)
OH2	0.42 (0.01)	0.43 (0.02)	0.42 (0.01)	0.43 (0.02)	0.42 (0.02)	0.42 (0.01)
C3	**0.34 (0.1)**	**0.42 (0.09)**	**0.24 (0.09)**	**0.20 (0.08)**	**0.39 (0.10)**	**0.30 (0.12)**
O3	−0.73 (0.03)	−0.76 (0.04)	−0.71 (0.03)	−0.73 (0.03)	−0.74 (0.02)	−0.72 (0.03)
OH3	0.43 (0.01)	0.43 (0.02)	0.43 (0.02)	0.44 (0.03)	0.43 (0.02)	0.43 (0.02)
C4	**0.19 (0.05)**	**0.12 (0.05)**	**0.33 (0.1)**	**0.40 (0.1)**	**0.18 (0.05)**	**0.26 (0.11)**
O4	−0.49 (0.04)	−0.47 (0.04)	−0.49 (0.05)	−0.46 (0.04)	−0.45 (0.04)	−0.47 (0.05)
C5	**0.32 (0.03)**	**0.31 (0.04)**	**0.22 (0.05)**	**0.20 (0.05)**	**0.28 (0.04)**	**0.24 (0.04)**
O5	−0.72 (0.03)	−0.67 (0.02)	−0.67 (0.02)	−0.69 (0.03)	−0.70 (0.02)	−0.67 (0.03)
OH5	0.42 (0.03)	0.41 (0.02)	0.42 (0.02)	0.43 (0.03)	0.42 (0.02)	0.42 (0.02)

[a] Partial atomic charges for the ring-averaged procedure are shown in the last column; numbers in parentheses correspond to standard deviations. [b] Puckering angles, P, and amplitudes, ϕ_m, are calculated according to the Altona-Sundaralingam method.[21] [c] For the ring-averaged results, P^* and ϕ_m^* indicate the most probable values based on the distribution shown in Figure 6.

these studies are similar to recent work by Woods and Kirschner[38] in which a similar analysis of hydroxymethyl groups on pyranoside rings was carried out. The notable differences here are that ring conformation is addressed and a new charge calculation procedure had to be implemented to take into account the flexibility of the furanoside ring.

Methods

Simulations. We adopted the combined AMBER/GLYCAM force field for the simulations of **1**. All the MD simulations were carried out using the AMBER 9.0[35] suite of programs, and the electronic structure calculations were performed with Gaussian 03.[39]

Solution Simulations. A 200 ns MD simulation of **1** was performed in a box of 298 TIP3P[40] water molecules under NPT conditions. The total box size was (25.569 × 25.372 × 25.544) (Å). The temperature was set to 300 K and the pressure to 1 atm. A cutoff of 8 Å was set for nonbonded interactions. The SCNB and SCEE scaling parameters were both set to 1.0 in accordance with the GLYCAM approach. All simulations were carried out under NPT conditions, and the SHAKE[41] algorithm was used to constrain all hydrogen-containing bonds. Prior to production MD simulations, minimization of the waters was first performed, followed by minimization of the whole system, 100 ps of annealing and 150 ps of equilibration. Ewald summation was used to handle long-range electrostatics.

Gas-Phase Simulations. The temperature was set to 300 K. A cutoff of 18 Å was set for nonbonded interactions. The SCNB and SCEE scaling parameters were both set to 1.0 in accordance with the GLYCAM approach. The SHAKE[41] algorithm was used to constrain all hydrogen-containing bonds.

Atomic Charges. Two charge derivation procedures were considered. The first one is the ensemble average approach proposed by Woods and workers[42] and is referred to as the usual GLYCAM procedure. Following this procedure, crystallographic data[43] were employed for the input geometry of methyl α-D-arabinofuranoside, and an ab initio geometry optimization was then performed at the HF/6-31G* level of theory. Based on the HF/6-31G* single point, the RESP[44] approach was used to obtain an initial set of restrained partial atomic charges. A relatively short MD simulation (10 ns) based on these charges and 100 conformations were selected from the resulting trajectory. The dihedral angles of the rotatable exocyclic moieties, such as hydroxyl groups, were then determined from the 100 snapshots and transferred to the quantum mechanics optimized geometry. Single point HF/6-31G* calculations were performed for these 100 new conformations. Partial atomic charges were obtained using the RESP approach for the 100 conformations, and the final charge of each atom was obtained as an average. The value of the RESP restraint weight was set to 0.01, and fitting was performed on all of the atoms except the aliphatic hydrogen.[45] The second charge derivation procedure is an important result of the current report and is described in the following section.

Results and Discussion

Atomic Charges. We present in Table 1 the atomic charges obtained from the standard GLYCAM procedure for five different ring conformers of **1**, labeled A−E. It is clear from this data that the charges vary when one changes the ring conformation. While this variation is not large for all atoms, the effect is especially pronounced for atoms C3, C4, and C5. For example, for C3 the charges vary over the range 0.20−0.42. This variability will negatively impact the accuracy of the simulations, and, to remove the bias associated with the choice of a specific ring conformation, we developed a charge averaging procedure that accounts for the various furanoside ring conformations.

Ring-Averaged Charges. Our modification of the usual GLYCAM approach, which incorporates the effects of the ring flexibility, is now described. Two hundred conformations were selected from a 50 ns simulation, and a constrained ab initio geometry optimization (HF/6-31G*) was performed for each. During those constrained optimizations, the dihedral angles involving hydroxyl protons were held to the values obtained from the MD simulation. For each of

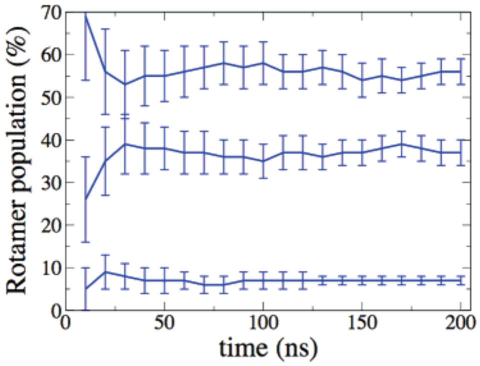


Figure 4. Convergence of the rotamer populations of **1**. Lines are a guide to the eye, and the *gg*, *gt*, and *tg* populations are given by the top, middle, and bottom lines, respectively.

the 200 new conformations, single point HF/6-31G* calculations were performed for the RESP fit. Note that the ring geometry and the dihedral angles involving hydroxyl protons are different in each of the 200 geometries. The same RESP approach as the one used in the usual GLYCAM procedure was then followed to obtain partial atomic charges.

The charges obtained from our new procedure, where charges are ensemble averaged over several exocyclic torsions *and* ring conformations, are presented in Table 1. We note that the new charges differ from those of the standard GLYCAM approach most notably for carbon atoms C3, C4, and C5. An average rmsd of the carbon atoms of the ring based on the 200 conformations used in the ring averaging was calculated, and a value of 0.09 with a fluctuation of 0.08 was obtained. This parameter is a convenient measure of the ring flexibility of the system. Along with the calculation of the rmsd, a correlation study between rmsd and puckering was carried out to quantify the magnitude of the rmsd in terms of puckering. In essence, this correlation study will indicate what change in ring puckering corresponds to a certain value of rmsd. However, this correlation study cannot be performed accurately on 200 conformations. It is necessary to consider many more conformations to get a statistically meaningful estimate. Therefore, we selected 100 000 conformations from the simulation based on our new ring-averaged atomic charges, whose results will be shown and discussed below. Based on that study, the current average rmsd of 0.09 corresponds to a change of about 60° in the puckering angle, *P*.

In the development of our ring average procedure, an alternate approach was attempted where one freezes not only the dihedral angles involving hydroxyl protons (as in our final average ring procedure) but also the dihedral angles of the ring (essentially fixing the ring puckering) in the geometry optimization of the 200 conformations selected from the simulation. In this way, the shape or puckering of the ring from the MD will be preserved, and our ring average will be more consistent with the simulation and, therefore, with the flexibility of the system. However, the geometry optimization of the 200 conformations with all these constraints did not converge. The conformations were over constrained, and all attempts to make them converge failed. The conformations extracted from the simulation seem to be very far from the ab initio minimum, and many constraints render convergence impossible.

Solution Simulations. Having determined the average atomic charges for **1**, we next set to establish the length of simulation required to achieve convergence. As a criteria for evaluating convergence we used the populations of rotamers about the C4−C5 bond. Shown in Figure 4 are the results of a convergence study of these rotamer populations in **1** as a function of simulation time. Charges obtained with the new ring-averaged procedure were used. From these results, it is clear that a 200 ns simulation is required to converge the populations of all the rotamers to reasonable uncertainties (a few units of percentage). Of particular note, simulations of less than 50 ns produced rotamer populations differing substantially from those present after 200 ns.

We next compared the C4−C5 rotamer populations obtained from the simulations with those derived from experimental results.[13] A histogram of the behavior of this torsion is shown in Figure 5. All three rotamers are populated, but the *tg* rotamer (180°) is visited infrequently. When the conformers from the three peaks in the histogram are integrated, it is possible to quantitate rotamer populations, which are presented in Table 2. In addition to the results based on our ring-averaged charge derivation procedure and the experimental values, the results of simulations based on the five charge sets of the standard (fixed ring) GLYCAM procedure are also presented. Clearly, the new ring-averaged charge calculation procedure leads to a good agreement with experiment, which is better than the fixed ring method. While both charge derivation approaches yield the correct ordering

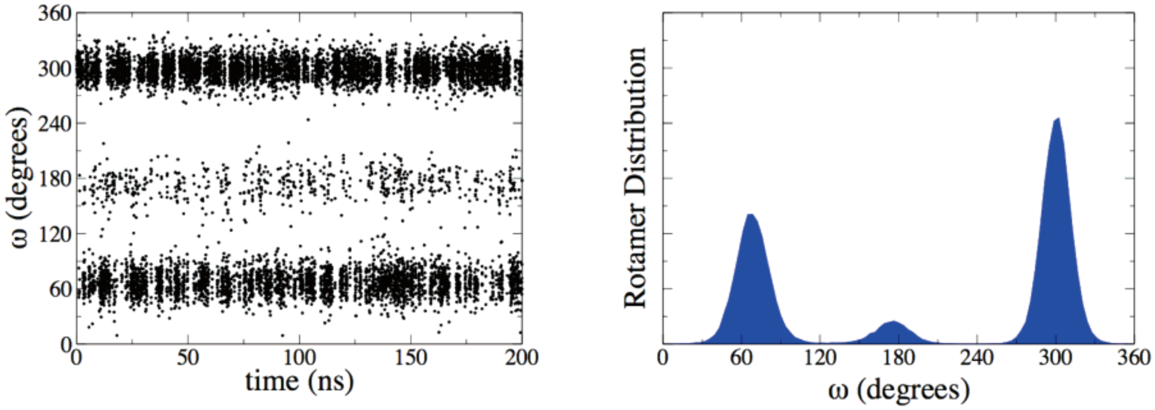


Figure 5. Time dependence of the C4−C5 torsion angle (left panel) and its associated distribution (right panel) for **1**.

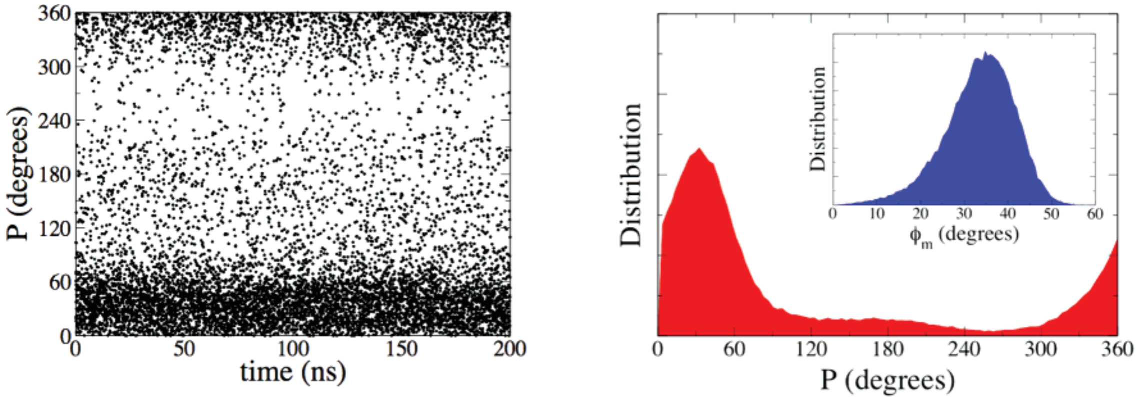


Figure 6. Time dependence of the Altona-Sundaralingam *P* angle (left panel) and its associated distribution (right panel) for **1**. The distribution of puckering amplitude, ϕ_m, is given in the inset of the right panel ($\phi_m^* = 35°$).

Table 2. Rotamer Populations of **1** Obtained Using the Various Approaches

rotamer population (%)	*gt*	*tg*	*gg*
experiment[13]	38	14	48
ring average charges	37(3)	7(1)	56(3)
fixed ring charges A	29(2)	8(1)	63(3)
fixed ring charges B	29(2)	8(1)	63(3)
fixed ring charges C	39(3)	7(1)	54(3)
fixed ring charges D	33(3)	8(1)	59(3)
fixed ring charges E	27(2)	8(2)	65(3)
gas phase	7(1)	40(3)	53(3)

of the rotamer populations, the results based on the usual GLYCAM approach can sometime lead to a worse agreement with experiment because of the intrinsic ring bias of that procedure. These results validate the ring-averaging method for obtaining charges in these flexible rings, and, encouraged by these results, we considered other ring parameters in **1**, in particular *P* and ϕ_m.

Figure 6 contains the variation in *P*, which describes ring puckering; the inset shows the variation in puckering amplitude, ϕ_m. The distribution in ϕ_m is centered about 35°, which corresponds well to earlier ab initio, density functional theory, and molecular calculations[15−19] on **1** as well as to the puckering amplitude of this molecule in the crystal structure.[43] With regard to *P*, conformations with values in the northern hemisphere of the pseudorotational itinerary (Figure 2) are clearly favored although a small fraction of the conformers are also present in the southern hemisphere. The area of conformational space centered about $P = 45°$ corresponds well to the N conformer determined for **1**[14] using the PSEUROT[22] procedure, which identified two conformers: a N conformer at $P = 44°$ (39%) and an S conformer at $P = 123°$ (61%). However, while there is good agreement with the identify of the N conformer, the conformer populations obtained from the simulation do not correspond well with experiment nor with previous ab initio and density functional theory calculations on **1**.[15−19] Indeed, the distribution shown in Figure 7 suggests that a while a small population of the S conformer (centered around $P = 180$) is present, the equilibrium is heavily biased to the N conformer. These results suggest that the two-state model inherent in the PSEUROT approach may not be valid for **1**.

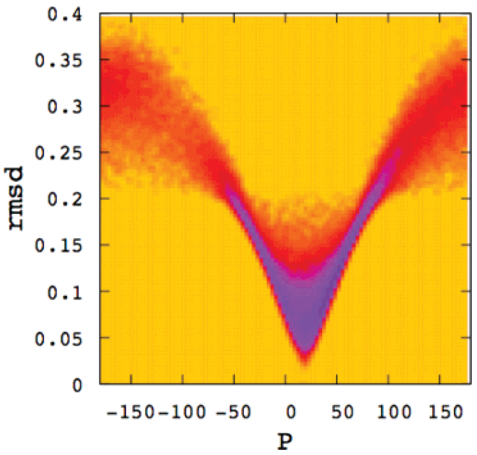


Figure 7. Joint probability distribution of the puckering angle (in degrees), *P*, and the rmsd (in Å) of the ring carbon atoms.

Figure 7 illustrates the correlation study mentioned earlier where we calculate the joint probability distribution of the puckering angle, *P*, and the rmsd of the ring atoms. The graph shows that a change of 180° in ring puckering, which is the maximum possible, represents a variation of approximately 0.25 in rmsd. The figure also reveals that an rmsd value of 0.09 as obtained in the ring-averaged charge derivation procedure of the preceding section corresponds to a 60° change in the puckering angle, *P*. If the fluctuation magnitude of 0.08 is taken into account, the change in ring puckering will be more than 100°. Obviously, this result lends weight to our modification to the standard GLYCAM procedure to derive the set of atomic charges. The current solvated molecular system is very flexible, and the charge derivation cannot be based on only one ring but has to be based on an average over numerous rings to represent all the conformations accessible to the system.

Gas-Phase Simulations. Although we anticipated that the inclusion of explicit water molecules to simulate solvent effects would be essential to obtain results consistent with experiment, as a test of this we performed a simulation of **1** in the gas phase. We present in Figure 8 the analysis of the C4−C5 torsion angle and, in Figure 9, pseudorotation behavior in the gas phase. As expected, these gas-phase results differ from those obtained with explicit solvent inclusion. This is presumably due, in large part, to the fact that in the

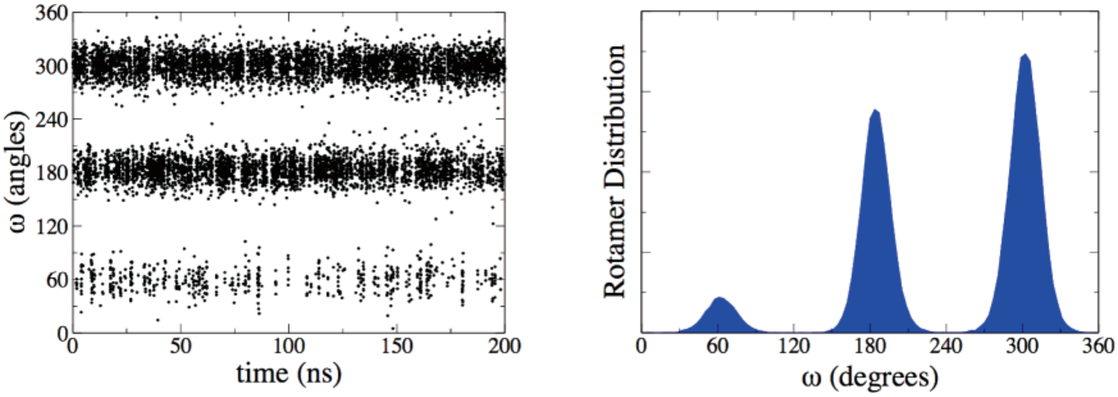


Figure 8. Time dependence of the C4−C5 torsion angle (left panel) and its associated distribution (right panel) for **1** in the gas phase.

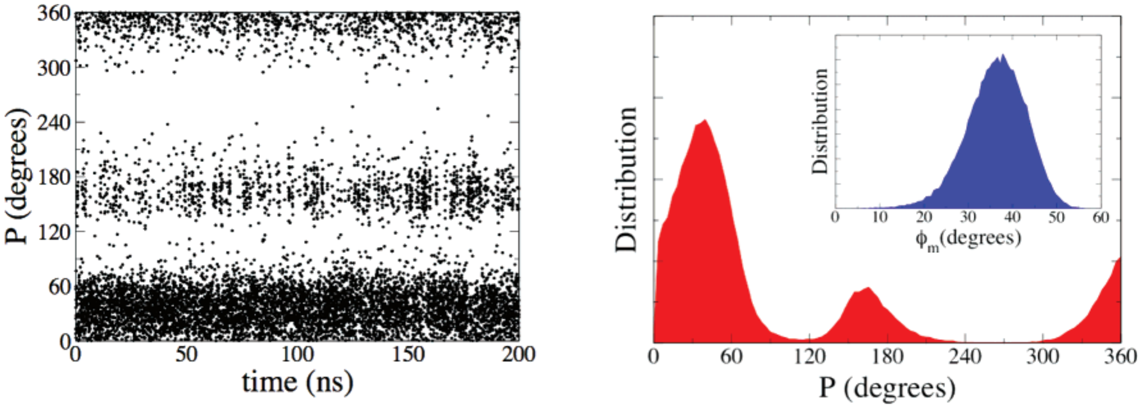


Figure 9. Time dependence of the *P* angle (left panel) and its associated distribution (right panel) for **1** in the gas phase (P_N* = 38 and P_S* = 165). The distribution of puckering amplitude, ϕ_m, is given in the inset of the right panel (ϕ_m* = 38).

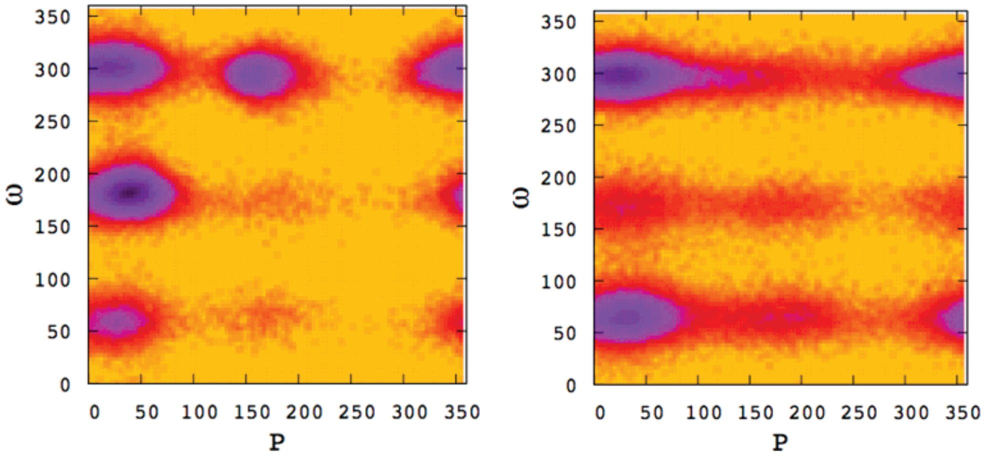


Figure 10. Joint probability distribution of the puckering angle, *P*, and the C4−C5 torsion for **1** in the gas (left panel) and solution (right panel) phases. The units of the angles *P* and ω are in degrees.

absence of water, the possibility of intermolecular hydrogen bond competition with the solvent is no longer possible.

We see from Figure 8 that the ordering of the rotamer populations is reversed compared to the solution and experimental cases. The population of the *tg* rotamer is now greatly enhanced at the expense of the *gt* rotamer. Figure 9 in turn reveals that the pseudorotation distribution now shows more distinct north (N) and south (S) populations. The most populated values of the two puckering states are P_N* = 38 and P_S* = 165 for the north and south regions, respectively, which agrees well with previous ab initio and density functional theory calculations on **1**.[15−19] This result differs significantly for the simulation done in the presence of water, where two distinct puckering states did not exist and instead a single region in the northern hemisphere of the pseudorotational itinerary was favored. As expected, these results underscore the importance of using an explicit solvent model to correctly describe solution behavior.

An ab initio and density functional theory study of several conformers of **1** in the gas phase[16] showed high correlation between the rotamer and the ring puckering distributions. In other words, the rotamer population depends on the ring puckering and vice versa.

Motivated by this study, we carried out a correlation study between the C4−C5 torsion and the puckering angle. Figure 10 shows the joint probability distribution of the C4−C5 torsion and puckering angle, *P*, for both gas- and solution-phase simulations. The gas-phase results reveal the presence of north and south hemispheres of the pseudorotational wheel, and different trends of C4−C5 torsion distribution are obtained for each hemisphere. For example, conformations with *P* values between 0 and 50° (North) exhibit the trend in rotamers of *tg* > *gg* > *gt*, whereas for conformations with *P* values around 180° (South), the trend is *gg* > *tg* = *gt*. The favoring of the *gg* rotamer in the S conformers would be expected given the ability of conformers with this C4−C5 torsion to form transannular hydrogen bonds between OH2 and OH5. Similarly, in the N conformers, the *tg* rotamer is stabilized by hydrogen bonding between OH3 and OH5. There is therefore a marked correlation between C4−C5 torsion and ring puckering in the gas phase, as concluded from an earlier ab initio study[16] although the trends in rotamers for the respective values of ring puckering do not coincide. The ab initio study shows *gg* > *gt* > *tg* for $P \approx 30$ and *gg* > *tg* > *gt* for $P \approx 180$. These differences may arise from the fact that in the ab initio study a full sampling of conformational space was not undertaken. Instead the energy-minimized structures were obtained by full optimization of a family of 30 ring-constrained conformers[15] that had been partially optimized to probe the effect of ring conformation on various molecular parameters, e.g., bond lengths and bond angles. In solution, this strong correlation between the C4−C5 rotamer and the furanose ring conformation is not observed. As seen in Figure 10, the north hemisphere of the pseudorotational wheel is mostly populated, regardless of the C4−C5 rotamer. We propose that the effect is due to the lack of intramolecular hydrogen bonding in the solution simulations.

Conclusions

In this paper, we have shown that the AMBER/GLYCAM model is applicable to furanoside rings, specifically methyl α-D-arabinofuranoside, **1**, provided that precautions are taken to account for the inherent flexibility of these five-membered rings. In particular, it is critical to use averaged atomic charges obtained from a large number of conformations (200). This approach leads to less charge variability and, in turn, more reproducible results. The usual GLYCAM procedure, in which a single ring conformer is used to derive atomic charges, appears to be valid for the more rigid pyranoside rings but not the conformationally mobile furanosides. Furthermore, long simulation times (200 ns) are required for convergence. Simulations in which these approaches were implemented showed good agreement with rotamer populations about the C4−C5 bond and the puckering amplitude of the ring (ϕ_m) as determined from NMR spectroscopic data.[17] In contrast, the results of the simulations in water demonstrated a single low-energy region of conformational space thus suggesting that the two-state model most frequently used to describe furanose ring conformation[17,22] may not be valid for **1**. In the gas-phase simulations, results consistent with the two-state model and earlier ab initio and density-functional theory calculations[15−19] were found.

The differences between ring conformer populations in the gas and aqueous phases are noteworthy and, while not necessarily unexpected, underscore both the profound influence of water on these flexible rings as well as the potential danger of consistently applying the two-state model in the conformational analysis of furanose moieties. In light of the present success of this approach to model these flexible rings, future studies will involve the extension of this method to the study of other commonly occurring furanoside monosaccharides (e.g., β-D-arabinofuranoside and β-D-galactofuranoside) as well as more complex oligomeric and polymeric structures related to mycobacterial arabinogalactan and lipoarabinomannan. Other issues such as the role of the water model or polarization will also be explored in forthcoming work.

Acknowledgment. The authors acknowledge the Natural Sciences and Engineering Research Council of Canada and the Alberta Ingenuity Centre for Carbohydrate Science for funding, and the WestGrid for access to computational resources.

References

(1) Saenger, W. *Principles of Nucleic Acid Structure*; Spring-Verlag: Berlin, 1988.

(2) Lowary, T. L. *Curr. Opin. Chem. Biol.* **2003**, *7*, 749−756.

(3) Houseknecht, J. B.; Lowary, T. L. *Curr. Opin. Chem. Biol.* **2001**, *5*, 677−682.

(4) Ryder, M. H.; Tate, M. E.; Jones, G. P. *J. Biol. Chem.* **1984**, *259*, 9704−9710.

(5) Komatsu, K.; Shigemori, H.; Kobayashi, J. *J. Org. Chem.* **2001**, *66*, 6189−6192.

(6) Gallego, J.; Varani, G. *Acc. Chem. Res.* **2001**, *34*, 836−843.

(7) Brennan, P. J.; Nikaido, H. *Annu. Rev. Biochem.* **1995**, *64*, 29−63.

(8) Crick, D. C.; Mahaptra, S.; Brennan, P. J. *Glycobiology* **2001**, *11*, 107R−118R.

(9) Briken, V.; Porcelli, S. A.; Besra, G. S.; Kremer, L. *Mol. Microbiol.* **2004**, *53*, 391−403.

(10) Minnikin, D. E. In *The Biology of the Mycobacteria*; Ratledge, C., Standford, J. L., Eds.; Academic: London, 1982; Vol. 1, pp 95−184.

(11) Angyal, S. J. *Adv. Carbohydr. Chem. Biochem.* **1984**, *42*, 15−68.

(12) Connell, N. D.; Nikaido, H. In *Tuberculosis: Pathogenesis, Protection and Control*; Bloom, B. R., Ed.; American Society for Microbiology: Washington, DC, 1994; pp 333−352.

(13) D'Souza, F. W.; Ayers, J. D.; McCarren, P. R.; Lowary, T. L. *J. Am. Chem. Soc.* **2000**, *122*, 1251−1260.

(14) Houseknecht, J. B.; Altona, C.; Hadad, C. M.; Lowary, T. L. *J. Org. Chem.* **2002**, *67*, 4647−4651.

(15) Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Am. Chem. Soc.* **1999**, *121*, 9682−9692.

(16) McCarren, P. R.; Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2001**, *105*, 5911−5922.

(17) Houseknecht, J. B.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2003**, *107*, 5763−5777.

(18) Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Org. Chem.* **2000**, *65*, 4954−4963.

(19) Houseknecht, J. B.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2003**, *107*, 372−378.

(20) Houseknecht, J. B.; McCarren, P. R.; Lowary, T. L.; Hadad, C. M. *J. Am. Chem. Soc.* **2001**, *123*, 8811−8824.

(21) Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* **1972**, *94*, 8205−8212.

(22) Deleeuw, F.; Altona, C. *J. Comput. Chem.* **1983**, *4*, 428−437.

(23) Lemieux, R. U.; Koto, S. *Tetrahedron* **1974**, *30*, 1933−1944.

(24) Wolfe, S. *Acc. Chem. Res.* **1972**, *5*, 102−111.

(25) Devries, N. K.; Buck, H. M. *Carbohydr. Res.* **1987**, *165*, 1−16.

(26) Bock, K.; Duus, J. O. *J. Carbohydr. Chem.* **1994**, *13*, 513−543.

(27) Tvaroska, I.; Carver, J. P. *J. Phys. Chem. B* **1997**, *101*, 2992−2999.

(28) Cros, S.; Dupenhoat, C. H.; Perez, S.; Imberty, A. *Carbohydr. Res.* **1993**, *248*, 81−93.

(29) Cros, S.; Imberty, A.; Bouchemal, N.; Dupenhoat, C. H.; Perez, S. *Biopolymers* **1994**, *34*, 1433−1447.

(30) Dowd, M. K.; French, A. D.; Reilly, P. J. *J. Carbohydr. Chem.* **2000**, *19*, 1091−1114.

(31) Mazeau, K.; Perez, S. *Carbohydr. Res.* **1998**, *311*, 203−217.

(32) French, A. D.; Dowd, M. K. *J. Comput. Chem.* **1994**, *15*, 561−570.

(33) French, A. D.; Dowd, M. K.; Reilly, P. J. *J. Mol. Struct. THEOCHEM* **1997**, *395*, 271−287.

(34) French, A. D.; Tran, V. *Biopolymers* **1990**, *29*, 1599−1611.

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

(36) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, B. *J. Phys. Chem.* **1995**, *99*, 3832−3846.

(37) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2007**, in press.

(38) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541−10545.

(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, Inc.: Wallingford, CT, 2004.

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

(41) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(42) Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22*, 1125−1137.

(43) Evdokimov, A. G.; Kalb, A. J.; Koetzle, T. F.; Klooster, W. T.; Martin, J. M. L. *J. Phys. Chem. A* **1999**, *103*, 744−753.

(44) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(45) Woods, R. J.; Chappelle, R. *J. Mol. Struct.* **2000**, *527*, 149−156.

CT700284R

JCTC Journal of Chemical Theory and Computation

Stability of the Free and Bound Microstates of a Mobile Loop of α-Amylase Obtained from the Absolute Entropy and Free Energy

Srinath Cheluvaraja and Hagai Meirovitch*

Department of Computational Biology, University of Pittsburgh School of Medicine, 3059 BST3, Pittsburgh, Pennsylvania 15260

Received May 16, 2007

Abstract: The hypothetical scanning molecular dynamics (HSMD) method is a relatively new technique for calculating the *absolute* entropy, S, and free energy, F, from a given sample generated by any simulation procedure. Thus, each sample conformation, i, is reconstructed by calculating transition probabilities that their product leads to the probability of i, hence to the entropy. HSMD is an exact method where all interactions are considered, and the only approximation is due to insufficient sampling. In previous studies HSMD (and HS Monte Carlo − HSMC) has been applied very successfully to liquid argon, TIP3P water, self-avoiding walks, and peptides in a α-helix, extended, and hairpin microstates. In this paper HSMD is developed further as applied to the flexible 7-residue surface loop, 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser) of the enzyme porcine pancreatic α-amylase. We are mainly interested in entropy and free energy differences $\Delta S = S_{\text{free}} - S_{\text{bound}}$ (and $\Delta F = F_{\text{free}} - F_{\text{bound}}$) between the free and bound microstates of the loop, which are obtained from two *separate* MD samples of these microstates without the need to carry out thermodynamic integration. As for peptides, we find that relatively large systematic errors in S_{free} and S_{bound} (and F_{free} and F_{bound}) are cancelled in ΔS (ΔF) which is thus obtained efficiently with high accuracy, i.e., with a statistical error of 0.1−0.2 kcal/mol (T=300 K) using the AMBER force field and AMBER with the implicit solvation GB/SA. We provide theoretical arguments in support of this cancellation, discuss in detail the problems involved in the computational definition of a microstate in conformational space, suggest potential ways for enhancing efficiency further, and describe the next development where explicit water will replace implicit solvation.

I. Introduction

I.1. The Role of Free Energy in Structural Biology. The theoretical/computational treatment of peptides, proteins, and other biological macromolecules is extremely difficult due to long-range interactions and their rugged potential energy surface, $E(\mathbf{x})$ ($\mathbf{x}$ is the $3N$-dimensional vector of the Cartesian coordinates of the molecule's N atoms). More specifically, this surface is "decorated" by a tremendous number of localized wells and "wider" ones, defined over regions, Ω_m (called microstates)—each consisting of many localized wells (an example for a microstate is the α-helical region of a peptide, see further discussion in sections II.3, II.11, and II.12). A microstate Ω_m, which typically constitutes only a tiny part of the entire conformational space Ω, can be represented by a sample (trajectory) generated by a *local* molecular dynamics (MD)[1,2] simulation starting from a structure that belongs to Ω_m. MD studies have shown that a molecule will visit a localized well only for a very short time [several femtoseconds (fs)] while staying for a much longer time within a microstate,[3,4] meaning that the microstates are of a greater physical significance than the localized wells.

* Corresponding author phone: (412)648-3338; e-mail: hagaim@pitt.edu.

10.1021/ct700116n CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/01/2007

A central aim of computational structural biology is to identify the most stable microstates, i.e., those with the largest *conformational* partition function Z_m (or equivalently with the lowest Helmholtz free energy, F_m)

$$F_m = -k_B T \ln Z_m = -k_B T \ln \int_m \exp[-E(\mathbf{x})/k_B T] d\mathbf{x} \quad (1)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and the integration is carried out over the limited microstate Ω_m, rather than over Ω (for simplicity, we shall denote in most cases a microstate Ω_m by m). Thus, the protein folding problem is the notoriously difficult task of identifying the microstate with the global minimum F_m, which practically might be achieved by two challenging stages: (1) identifying an initial set of microstates with expected high stability (e.g., based on an energetic criterion) and (2) calculating their relative populations, $p_m/p_n = Z_m/Z_n$ [$p_m = \exp[-F_m/k_B T]/Z$, where Z is the (cancelled out) partition function of the entire conformational space, Ω], which leads to minimum F_m

$$p_m/p_n = Z_m/Z_n = \exp - [\Delta F_{\boldsymbol{mn}}/k_B T] \quad (2)$$

where $\Delta F_{\boldsymbol{mn}} = F_{\boldsymbol{m}} - F_{\boldsymbol{n}}$.

Calculation of relative populations is also required in problems which are less challenging than protein folding, i.e., in cases of *intermediate flexibility*, where a flexible protein segment (e.g., a side chain or a surface loop), a cyclic peptide, or a ligand bound to an enzyme populates significantly several microstates in thermodynamic equilibrium. It is of interest to know whether the conformational change adopted by a loop (a side chain, ligand, etc.) upon binding has been induced by the other protein (induced fit[5,6]) or alternatively the free loop already interconverts among different microstates where one of them is selected upon binding (selected fit[7]). This analysis requires calculating p_m values, which are also needed for a correct analysis of NMR and X-ray data of flexible macromolecules.[8−11] Calculation of F is essential in many other biological processes. Thus, F determines the binding affinities of protein−protein interactions, it is an important factor in enzymatic reactions, electron transfer, and ion transport through membranes, and it leads to the solubilities of small molecules.

I.2. The Difficulty in Calculating the Free Energy. It should first be pointed out that the *absolute* Helmholtz free energy is $F_m = E_m - TS_m$ where S_m is the absolute entropy. Monte Carlo (MC)[12] and MD[1,2] are dynamic methods, which enable one to generate samples of system configurations, $\mathbf{x}$ distributed according to their Boltzmann probability density, $\rho^B(\mathbf{x})$

$$\rho^B(\mathbf{x}) = \exp[-E(\mathbf{x})/k_B T]/Z_m \quad (3)$$

(Z_m is defined over m or the entire conformational space, Ω). With both methods it is straightforward to estimate ensemble averages of quantities that are measured directly from $\mathbf{x}$, such as $E(\mathbf{x})$. On the other hand, to estimate the entropy (defined up to an additive constant)

$$S_m = -k_B \int_m \rho^B(\mathbf{x}) \ln \rho^B(\mathbf{x}) d\mathbf{x} \quad (4)$$

one has to calculate the practically unknown *value* of $\ln \rho^B(\mathbf{x})$ [$\rho^B(\mathbf{x})$ depends not only on $\mathbf{x}$ but also on the entire microstate through Z_m, where Z_m is extremely difficult to calculate directly from the sample]. Thus, the difficulty in calculating F_m stems from the difficulty in calculating S_m. In most cases, however, one is interested in free energy differences ΔF_{mn}, which are somewhat easier to obtain than F_m and F_n themselves.[13−19]

I.3. Calculation of ΔF_{mn} by the Counting Method and Thermodynamic Integration. As said above, even calculation of relative populations is nontrivial. A straightforward way to estimate $p_n/p_m = \exp - [\Delta F_{mn}/k_B T]$ is by a *counting method*, i.e., from a long MD or MC simulation that "covers" both microstates. Thus, $\Delta F_{\boldsymbol{mn}} = -k_B T \ln[(\#m)/(\#n)]$, where $\#m$ ($\#n$) is the population, i.e., the number of times the molecule visited microstate m (n) during the simulation. However, because of high-energy barriers, the transition between microstates at room temperature might require long times, nanoseconds or more even for side-chain rotamers, meaning that reliable sampling of $\#m$ ($\#n$) might become prohibitive. This problem can be alleviated by applying enhanced sampling techniques such as replica exchange[20] or multicanonical methods;[21,22] however, the conformational search capability of these methods is also limited, and microstates of interest might be visited poorly or will not be visited at all. The common analysis is based on projecting MD (MC) trajectories onto a small number of coordinates using principal component analysis or calculating the populations along one or two physically significant reaction coordinates.[23,24]

Differences, ΔF_{mn}, are commonly calculated by thermodynamic integration (TI) over physical quantities such as the energy, temperature, and the specific heat[25,26] as well as nonphysical parameters[13−19,27−34] (free energy perturbation methods and umbrella and histogram analysis methods[35−37] are also included in this category, see ref 19 and references cited therein). This is a robust and highly versatile approach, which is used successfully for calculating the difference in the free energy of binding of two ligands to the active site of an enzyme. However, if the structural variance of m and n is large, then the integration from m to n becomes difficult and in many cases unfeasible. Furthermore, because MC (MD) simulations constitute models for dynamical processes, one would seek to calculate changes in F and S during a relaxation process, by assuming local equilibrium in certain parts along the trajectory; a classic example is simulation of protein folding.[38] Such information cannot be obtained by TI, and it is thus desirable to develop methods that estimate S and F directly from a given trajectory.

I.4. Calculation of the Absolute Entropy. The problems in calculating ΔF_{mn} mentioned above could be remedied to a large extent by developing methods for calculating the *absolute* F_m from a given sample. This would enable one to carry out (only) two *separate* MD simulations of microstates m and n, calculating directly the absolute $F_{\boldsymbol{m}}$ and $F_{\boldsymbol{n}}$ and their difference $\Delta F_{mn} = F_m - F_n$, where the TI process or the long runs needed in the counting method are avoided.

A commonly used approach for estimating the absolute S is based on the harmonic approximation and was introduced to biomolecules by Gō and Scheraga.[39,40] They obtained $S = -(k_B/2)\ln[\text{Det(Hessian)}]$, where Hessian is the matrix of second derivatives of the force field around an energy minimized structure; the quantum mechanical version was applied later for peptides.[41] An important development has been the introduction of the quasi-harmonic (QH) method by Karplus and Kushick,[42] where the Boltzmann probability density of structures defining a microstate is approximated by a multivariate Gaussian. Thus,

$$S_m^{\text{QH}} = (k_B/2)\{N + \ln[(2\pi)^N \,\text{Det}(\sigma)]\} \qquad (5)$$

where the covariance matrix, σ, is obtained from a local MD (MC) sample, and N is (usually) the number of internal coordinates. Clearly, S^{QH} constitutes an upper bound for S since correlations higher than quadratic are neglected; also, anharmonic contributions are ignored, and QH is not suitable for diffusive systems such as water. While QH has been used extensively during the years, a systematic study of its performance has been carried out only recently by Gilson's group[43] who have found that the performance of QH deteriorates significantly in Cartesian coordinates and when applied to more than one microstate.[19]

The absolute F can also be obtained with TI provided that a reference state R with known F_R is available and an efficient integration path $R \rightarrow m$ can be defined. A classic example is the calculation of F of liquid argon or water by integrating the free energy from an ideal gas reference state. However, for nonhomogeneous systems such integration might not be trivial, and in models of peptides and proteins defining adequate reference states is a problem. Differences in free energy can be obtained by Bennett's method and techniques that are derivatives of Bennett's method (for a more complete discussion about methods for calculating the absolute entropy see ref 19).

I.5. Our Methods for Calculating the Absolute *S*. Another approach for calculating the absolute S (F) has been suggested by Meirovitch and has been implemented in two *approximate* techniques of general applicability (i.e., they are not restricted to harmonic conditions), the local states (LS)[44−46] and the hypothetical scanning (HS)[47−49] methods. With both methods each conformation i of an MC(MD) sample is *reconstructed* step-by-step (from nothing) using transition probabilities (TPs), where their product leads to an approximation for the correct Boltzmann probability (eq 3) from which various free energy functionals (e.g., upper and lower bounds) can be defined. Recently, the deterministic approximate calculation of TP(HS) was replaced by a stochastic calculation carried out by MC(MD) simulations, where *all* interactions are taken into account, and from this respect the method [called HSMC(D)] can be viewed as exact;[50] the only approximation involved is due to insufficient MC(MD) sampling. HSMC(D) has unique features: it provides *rigorous* lower and upper bounds for F, which enable one to determine the accuracy from HSMC(D) results alone without the need to know the correct answer. Furthermore, F can be obtained from a very small sample and even from *any single* conformation. HSMC results, which agree within error bars with TI results, were obtained for liquid argon, TIP3P water,[50,51] self-avoiding walks on a square lattice,[52] and peptides.[53,54] Very recently HSMD has been extended to peptides with side chains simulated by MD.[55] We have found that reliable results for *differences*, ΔS_{mn} and ΔF_{mn}, can be obtained with considerable efficiency, ~100 times faster (in term of computer time) than with MC. These results obtained for decaglycine and $NH_2(Val)_2(Gly)_6$-$(Val)_2CONH_2$ are very encouraging, suggesting that HSMD might become a highly efficient tool for calculating ΔF_{mn} (our main interest) also for more complex systems such as loops.

I.6. A Mobile Loop in Porcine Pancreatic α-Amylase. In this paper we develop HSMD further by applying it to a flexible surface loop of the enzyme porcine pancreatic α-amylase (PPA). α-Amylases (α-1,4-glucan-4-glucanohydrolases, EC 3.2.1.1) are widespread in all three domains of life: Archaea, Bacteria, and Eucarya. These enzymes catalyze the hydrolysis of internal glycosidic bonds in starch and related poly- and oligosaccharides. α-Amylases play a central role in carbohydrate metabolism of microorganisms, plants, and animals. Furthermore, they are widely used in the food and starch processing industry. Many of the enzymatic studies have been carried out with PPA, which serves as a model system.

PPA is a single polypeptide chain of 496 amino acid residues[56−59] consisting of three structural domains, domain A (residues 1−99, 170−404), domain B (residues 100−169), and domain C (residues 405−496). Domain A adopts a $(\beta/\alpha)_8$ barrel structure and contains the three catalytic residues Asp197, Glu233, and Asp300. Domain B occurs as an excursion from domain A and is the structurally least ordered of the three domains; it contains one calcium-binding site. Domain C forms an all-β structure and seems to be an independent domain with unknown function.[56] The active site and the possible roles of associated ions have been well characterized from the crystal structures of several amylases. A deep cleft in domain A is accepted to be the substrate-binding site.[56−62] An essential chloride ion and a calcium ion are located closer to this V-shaped depression and have been suggested to enhance the catalytic activity.[62−65]

While substantial evidence is available for the role of catalytic residues in amylases, very few studies have been carried out on the role of loops surrounding the active site that interact with the substrate. In the crystal structures of the free protein (PPA I[56] and II[59] which differ by two residues) loop 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser) has larger B-factors than the average B-factors of the atoms in the protein. However, in the crystal structures of PPA I complexed with acarbose[57] and PPA II complexed with V-1532[59] the B-factors of this loop are close to the average value in the protein where the loop has moved toward the active site. The maximum main-chain movement is ~5 Å at His305, which approaches the inhibitor from the solvent side to make a hydrogen bond with a glucose residue. The outcome of this movement is an apparent closure of the surface edge of the cleft.[57] Subsequently, several hypotheses have been put forward with respect to the function of the mobile loop in α-amylases, such as providing assistance in holding the glucose residues in a favorable orientation during catalysis,[57] or assisting in the transition state,[66] or inducing a trap-release mechanism of substrate and products.[67]

I.7. Extension of HSMD for Loops. This work constitutes the first step in extending HSMC(D) to surface loops in proteins, for which the above short mobile loop (with its small residues) serves as an ideal system. Two MD simulations, starting from the X-ray structures of the free and complexed PPA II, 1pif and 1pig, respectively,[59] will span the corresponding microstates, and the entropy and free energy will be calculated by HSMD. In this initial study the loop is modeled by the AMBER force field[68] alone (where solvation effects are not considered) and by the AMBER and the highly approximate GB/SA implicit solvent.[69] Therefore, the study is focused mainly on (technical) implementation issues of HSMD rather than on the role of the loop in the enzymatic function of PPA; the latter subject will be discussed in future studies where explicit water will be introduced. Still, the present study might indicate whether the transition of the loop to the bound microstate constitutes a selected fit, i.e., whether this microstate is reachable in the free protein. We also discuss in detail various theoretical aspects of HSMD elaborating in particular on the problematic definition of a microstate.

II. Theory and Methodology

II.1. The Protein and Loop Studied. As was pointed out in section I.6 we study the loop of $N = 7$ residues, 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser), of PPA in two microstates related to the free and bound loop structures. The starting point is the available crystal structures of PPA II, 1pif and 1pig,[59] respectively. Because the structures of these proteins are almost identical, we have chosen to carry out the calculations with the 1pif structure, where the loop structure of 1pig is attached to the 1pif structure by superimposing the structure of 1pig on that of 1pif (the ligand was discarded); this would enable one to study the stability of the bound microstate of the loop in the structure of the free protein, as discussed in the previous section I.7. PPA is a relatively large protein (496 residues), and it would be computationally unfeasible to include all of its atoms in the calculations. Therefore, we consider only a template of 700 atoms (the same atoms for the bound and free structures) that are close to the loop where the rest of the protein's atoms are ignored. The construction of the template is described in detail in a previous publication.[51]

The loops are modeled in vacuum where the potential energy is defined solely by the AMBER96 force field[68] and in solution where the implicit solvation model, GB/SA,[69] is added to this force field. The His residue is protonated in the free and bound states. These systems are simulated by MD using the package TINKER,[70] where the loop is free to move while the template (of 700 atoms) is kept fixed in its X-ray coordinates and only the loop−loop and loop−template interactions are considered, i.e., they define the potential energy. However, HSMD (as well as LS and QH) is implemented naturally in internal coordinates; therefore, the simulated conformations should be transferred from Cartesians to the dihedral angles φ_i, ψ_i, and ω_i (i=1,N=7) and the bond angles $\theta_{i,l}$ (i=1,N, l=1,3) and the side-chain angles χ and the corresponding bond angles. For the present loop we consider two χ angles one of His and one of Ser, while the contribution of the side chain of Ala is ignored; also, because the side chains are much shorter than the backbone and are not restricted by the loop closure condition, the effect of their bond angles on entropy *differences* is expected to be small and is thus ignored (in the next section we argue that to a good approximation bond stretching can be ignored as well). For convenience, these angles (ordered along the backbone) are denoted by α_k, $k = 1,45 = K$.

II.2. Statistical Mechanics of a Loop in Internal Coordinates. The partition function Z_m (eq 1) of a loop is an integration of exp $-$ $[E(\mathbf{x})/k_BT]$ with respect to the loop's Cartesian coordinates, $\mathbf{x}$ over a microstate m. The change of the variables of integration from $\mathbf{x}$ to internal coordinates, α_k, $k = 1,K$, makes the integral dependent also on a Jacobian, J, which for a linear chain has been shown to be a simple function of the bond angles and bond lengths independent of the dihedral angles.[39,40,42] This transformation is applied under the assumption that the potentials of the bond lengths ("the hard variables") are strong, and, therefore, their average values can be assigned to J, which to a good approximation can be taken out of the integral (however, see a later discussion in this section). For the same reason one can carry out the integration over the bond lengths (assuming that they are not correlations with the α_k), and the remaining integral becomes a function of the K dihedral and bond angles (α_k)[39,40,42] and a Jacobian that depends only on the bond angles; the same discussion also holds for a loop. The partition function becomes

$$Z'_m = DZ_m = D \int_m \exp\{-E([\alpha_k])/k_BT\} d\alpha_1 ... d\alpha_K \quad (6)$$

where $[\alpha_k] = [\alpha_1,...\alpha_K]$. D is a product of the integral over the bond lengths and their Jacobian J. The Jacobian $[\Pi_j \sin(\theta_j)]$ of the bond angles, θ_j, that should appear under the integral is omitted for simplicity. We *assume* D to be the same (i.e., constant) for different microstates of the same loop, and, therefore, ln$D$ cancels and can be ignored in calculations of free energy and entropy *differences*. The Boltzmann probability density corresponding to Z_m (eq 6) is

$$\rho^B([\alpha_k]) = \exp\{-E([\alpha_k])/k_BT\}/Z_m \quad (7)$$

and the exact entropy S and exact free energy F (defined up to an additive constant) are

$$S_m = -k_B \int_m \rho^B([\alpha_k]) \ln \rho^B([\alpha_k]) d\alpha_1 .... \alpha_K \quad (8)$$

and

$$F_m = \int_m \rho^B([\alpha_k])\{E([\alpha_k]) + k_BT \ln \rho^B([\alpha_k])\}\, d\alpha_1 .... \alpha_K \quad (9)$$

It should be pointed out that the fluctuation of the *exact* F is zero,[71] because by substituting $\rho^B([\alpha_k])$ (eq 7) inside the curly brackets of eq 9 one obtains $E([\alpha_k]) + k_BT \ln \rho^B([\alpha_k]) = -kT \ln Z_m = F_m$, i.e. the expression in the curly brackets is constant and equal to F_m for any set $[\alpha_k]$ within m. This means that the free energy can be obtained from *any single* conformation if its Boltzmann probability density is known. However, the fluctuation of an approximate free energy (i.e., which is based on an approximate probability density) is finite, and it is expected to decrease as the approximation improves.[49,71−74] Using the HSMC(D) method, it is possible to estimate the free energy of the system from any single structure.

With MD the bond stretching energy is taken into account in eq 9 (and in free energy functionals defined later), while the corresponding entropy is ignored. The contribution of this energy to the free energy becomes an additive constant if one accepts the assumptions about the stretching energy and the corresponding Jacobian made prior to eq 6. This is a very good approximation; however, if the bond stretching entropy should be considered, we argue in section II.6 that it can be estimated *approximately* within the framework of HSMD by assuming that bond stretching is independent of the other interactions.

II.3. On the Practical Definition of a Microstate Thus far we have defined microstates in general terms and discussed various techniques for calculating their populations, p_m (or the ratios p_m/p_n); however, such calculations cannot be carried out without first establishing a *practical* definition of a microstate, which is not straightforward. Therefore, before discussing the theory further, we elaborate about this important issue that has been ignored to a large extent in the literature but has been given considerable thought by us over the course of years.[9,45,46,74−77] For simplicity we consider (for this discussion) an N-residue peptide in a helical microstate Ω_h with constant bond lengths and bond angles (ω_i=180°) meaning that its backbone conformation is solely defined by the angles, φ_i and ψ_i (i=1,N); in Ω_h these angles are expected to vary within relatively small ranges $\Delta\varphi_i$ and $\Delta\psi_i$ around $\varphi_i = -60°$ and $\psi_i = -50°$ (we ignore for a moment the side chains). However, if N is not too small, the correct limits of Ω_h in terms of $[\varphi_i,\psi_i]$ are unknown even for this simplified model because the strongly correlated angles define a complicated narrow "pipe" within the region, $\Delta\varphi_1\mathrm{x}\Delta\psi_1\mathrm{x}\Delta\varphi_2\mathrm{x}\Delta\psi_2 \cdots\cdots \Delta\varphi_N\mathrm{x}\Delta\psi_N$. Obviously, these correlations are taken into account by an exact simulation method, and, thus, in practice, Ω_h can be defined (or more correctly, represented) by a *local* MC (MD) sample of conformations initiated from an α-helical structure (as mentioned in section I.1).

However, this definition should be used with caution. Thus, a short simulation will span only a small part of Ω_h, and this part will grow constantly as the simulation continues; correspondingly, the calculated average potential energy, E_h, and the entropy, S_h (obtained by any method), will both increase, and the free energy, F_h, is expected to change as well. As the simulation time is increased further, side-chain dihedrals will "jump" to different rotamers, which according to our definition should also be included within Ω_h; for a long enough simulation the peptide is expected to "leave" the α-helical region moving to a different microstate. Thus, *in practice*, the microstate size and the corresponding thermodynamic quantities depend on the simulation time. In some cases, one can better define Ω_h by discarding structures with dihedral angles beyond predefined $\Delta\varphi_i$ and $\Delta\psi_i$ values or structures that do not satisfy a certain number of hydrogen bonds; one can also apply energetic restraints where their bias should later be removed. However, these restrictions are somewhat arbitrary and are difficult to apply for calculating the differences ΔF_{mn} and ΔS_{mn} between microstates m and n, *which is our main interest*. Therefore, in practice there is always some arbitrariness in the definition of a microstate, which affects the calculated averages. This arbitrariness is severe with some methods and can be controlled (minimized) by others, as is discussed in sections II.9 and II.10.

II.4. Exact Scanning Procedure. The HS, LS, and HSMC(D) methods are based on the ideas of the *exact* scanning method, which is a step-by-step construction procedure for a peptide.[78,79] For simplicity this construction is described for an N-residue polyglycine molecule (with dihedral and bond angles denoted α_k, $1 \le \alpha_k \le 6N = K$) in a microstate m. Thus, starting from nothing, a conformation of this molecule is built by defining the angles α_k step-by-step with transition probabilities (TPs) and adding the related atoms;[79] for example, the angle φ determines the coordinates of the two hydrogens connected to C^α, while the bond angle $\mathrm{N}{-}\mathrm{C}^\alpha{-}\mathrm{C}'$ determines the position of C′. Thus, at step k, $k-1$ angles $\alpha_1, \cdots, \alpha_{k-1}$ have already been determined; these angles and the related structure (the past) are kept constant, and α_k is defined with the *exact* TP density $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$

$$\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) = Z_\mathrm{future}(\alpha_k \cdots \alpha_1)/[Z_\mathrm{future}(\alpha_{k-1} \cdots \alpha_1)] \qquad (10)$$

where $Z_\mathrm{future}(\alpha_k \cdots \alpha_1)$ is a future partition function defined over m by integrating over the future conformations defined by $\alpha_{k+1} \cdots d\alpha_K$ (within m) where the past angles, $\alpha_1 \cdots \alpha_k$ (and their corresponding atoms), are held fixed

$$Z_\mathrm{future}(\alpha_k \cdots \alpha_1) = \int_m \exp - [(E(\alpha_K, \cdots, \alpha_1))/k_\mathrm{B}T] d\alpha_{k+1} \cdots d\alpha_K \qquad (11)$$

For simplicity, from now on we shall omit in most cases the subscript m from the thermodynamic functions. The product of the TPs (eq 10) leads to the probability density of the entire conformation (eq 7)

$$\rho^\mathrm{B}(\alpha_K, \cdots, \alpha_1) = \prod_{k=1}^{K} \rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \qquad (12)$$

This construction procedure is not feasible for a large molecule because the scanning range grows exponentially and the limits of the microstate m are practically unknown, as discussed in section II.3 (for a practical use of this method see ref 79). However, the exact scanning method constitutes the basis for HS as well as for the much less restricted HSMC(D) and LS methods. The exact scanning method is applicable to a peptide (loop) with side chains,[55] where for a loop all the backbone future conformations should also satisfy the loop closure condition.

The *exact* scanning method is equivalent to any other exact simulation technique (in particular MC and MD) in the sense that large samples generated by such methods lead to the same averages and fluctuations. Therefore, one can assume that a given MC or MD sample has rather been generated by the exact scanning method, which enables one to reconstruct each conformation *i* by calculating the TP densities that *hypothetically* were used to create it step-by-step. With HSMC(D) (unlike HS) the *entire* future is considered in the reconstruction process, and in this respect HSMC(D) can be considered to be exact.

II.5. The HSMC(D) Method. The theory is described for HSMD (which is more efficient and practical than HSMC[55]) as applied (for simplicity) to an *N*-residue polyglycine molecule. One starts by generating an MD sample of microstate *m*; the conformations are then represented in terms of the dihedral and bond angles α_k, $1 \leq \alpha_k \leq 6N = K$, and the variability range $\Delta\alpha_k$ is calculated

$$\Delta\alpha_k = \alpha_k(\max) - \alpha_k(\min) \qquad (13)$$

where $\alpha_k(\max)$ and $\alpha_k(\min)$ are the maximum and minimum values of α_k found in the sample, respectively. $\Delta\alpha_k$, $\alpha_k(\max)$, and $\alpha_k(\min)$ enable one to verify that the sample spans correctly the microstate *m*.

As pointed out in section II.4, with our approach a sample conformation *i* is reconstructed step-by-step by calculating the TP density of each α_k (eq 10) from the future partition functions Z_{future} (eq 11). However, a deterministic integration of Z_{future} based on the *entire* future (within the limits of *m*) is difficult and becomes impractical for a large peptide where *m* is unknown (see section II.3). The idea of HSMD is to obtain the TPs (eq 10) by carrying out MD simulations of the future part of the chain rather than by evaluating the integrals defining Z_{future} (eq 11) in a deterministic way. Thus, at reconstruction step *k* of conformation *i* the TP density, $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$, is calculated from an MD sample of n_f conformations (generated in Cartesian coordinates), where the *entire* future of the peptide is moved (i.e., the atoms defined by $\alpha_k, \cdots, \alpha_K$) while the past (the atoms defined by $\alpha_1, \cdots, \alpha_{k-1}$) are kept fixed at their values in conformation *i*. A small segment (bin) $\delta\alpha_k$ is centered at $\alpha_k(i)$, and the number of visits of the future chain to this bin during the simulation, n_{visit}, is calculated; one obtains

$$\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \approx \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1) = n_{\text{visit}}/[n_f \delta\alpha_k] \qquad (14)$$

where $\rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ becomes exact for very large n_f ($n_f \rightarrow \infty$) and a very small bin ($\delta\alpha_k \rightarrow 0$). This means that in practice $\rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ will be somewhat approximate due to insufficient future sampling (finite n_f), a relatively large bin size $\delta\alpha_k$, an imperfect random number generator, etc. Because this TP is also applicable to HSMC, we denote it (and functions derived from it) with 'HS' (rather than 'HSMD'). Notice that with HSMD the future conformations generated by MD at each step *k* remain in general within the limits of *m*, which is represented by the analyzed MD sample. The corresponding probability density is

$$\rho^{\text{HS}}(\alpha_K, \cdots, \alpha_1) = \prod_{k=1}^{K} \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \qquad (15)$$

$\rho^{\text{HS}}([\alpha_k])$ defines approximate entropy and free energy functionals, S^{A} and F^{A}, which can be shown using Jensen's inequality to constitute *rigorous* upper and lower bounds, respectively[50]

$$S^{\text{A}} = -k_{\text{B}} \int_m \rho^{\text{B}}([\alpha_k]) \ln \rho^{\text{HS}}([\alpha_k]) d\alpha_1 \cdots \alpha_K \qquad (16)$$

$$F^{\text{A}} = \langle E \rangle - TS^{\text{A}} = \langle E \rangle + k_{\text{B}}T \int_m \rho^{\text{B}}([\alpha_k]) \ln \rho^{\text{HS}}([\alpha_k]) d\alpha_1 \cdots \alpha_K \qquad (17)$$

where $<E>$ is the Boltzmann average of the potential (force field) energy, estimated from the MD sample, and ρ^{B} (eq 7) is the Boltzmann probability density with which the sample has been generated. S^{A} is estimated from a Boltzmann sample of size *n* by the arithmetic average, $\overline{S^{\text{A}}}$

$$\overline{S^{\text{A}}} = \frac{1}{n} \sum_{t=1}^{n} \ln \rho_t^{\text{HS}} \qquad (18)$$

As discussed in section II.2, the fluctuation (standard deviation) of the correct free energy (eq 9) is zero, while the approximate F^{A} has finite fluctuation, σ_{A} (estimated by its arithmetic average, $\overline{\sigma_{\text{A}}}$), which is expected to decrease as the approximation improves (i.e., as n_f increases and/or $\delta\alpha_k$ decreases)[49,71−74]

$$\overline{\sigma_{\text{A}}} = \left[\frac{1}{n} \sum_{t=1}^{n} [\bar{\text{F}}^{\text{A}} - E_t - k_{\text{B}}T \ln \rho_t^{\text{HS}}]^2\right]^{1/2} \qquad (19)$$

While (for simplicity) in the theory above only a single angle is reconstructed at each step *k*, in practice a pair of angles is treated simultaneously, where each pair consists of a dihedral angle and its successive bond angle (e.g., φ and the bond angle N−C$^{\alpha}$−C′). Thus, at each step both α_k and α_{k+1} are considered, and n_{visit} is increased by 1 only if α_k and α_{k+1} are located within the limits of $\delta\alpha_k$ and $\delta\alpha_{k+1}$, respectively. The HSMD process described above for polyglycine is also applicable to a side chain and a loop, where the reconstruction process of the latter starts from the first residue (which is connected to one end of the template), and the future chains are always connected (by the force field) to the second end of the template. Clearly, the conformational freedom of the future chains decreases as step *k* increases.

It should be pointed out again that in the case of HSMD the dependence of F^{A} (eq 17) on the bond stretching energy is only through $<E>$, while this interaction is ignored in S^{A}. However, under the assumptions leading to eq 6 this is not expected to affect differences in free energy which are our main interest. Still, if one seeks to include the bond stretching entropy, one can use a transition probability density, $\rho(a_k)$, similar to eq 14 for the bond length a_k which corresponds to the pair of atoms *k* and *k*+1; considering the Jacobian, one obtains $\rho(a_k) \approx n_{\text{visit}}(a_k)/[n_f 3^{-1} \delta(a_k^3)]_,$ where δa_k is small compared to a_k and $n_{\text{visit}}(a_k)$ is the number of visits to a_k. In this *approximation* the bond stretching is independent of the other interactions and thus $\rho_{\text{TP}}^{\text{HS}} = \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)\rho(a_k)$. Both probability densities can be calculated simultaneously, which in practice would not increase computer time.

II.6. The Reconstruction Procedure with HSMD. The HSMD reconstruction procedure needs further discussions.

Thus, the MD simulation of the future chain at step k starts from the reconstructed conformation i, and every g fs the current conformation is considered, where the n_{init} initial considered conformations are discarded for equilibration. The next n_f (considered) future conformations are represented in internal coordinates, and their contribution to n_{visit} (eq 14) is calculated. An essential issue is how to guarantee an adequate coverage of microstate m, i.e., that the future chains will span its entire region (in particular the side-chain rotamers) while avoiding their "overflow" to neighbor microstates, conditions that will occur for a too small and a too large n_f, respectively. To be able to control the extent of coverage of m the following procedure has been applied: n_f has been divided into several (j) shorter repetitive procedures ("units"), each based on $n'_f < n_f$ conformations where $n_f = jn'_f$, and each unit starts from the reconstructed structure i with a different set of velocities followed by equilibration of size, n_{init}; obviously, one would seek to determine the minimal values for n'_f, j, and n_{init}, which would keep the future chains within m while allowing its adequate sampling. A similar procedure was first suggested by Brady and Karplus[80−82] within the framework of the QH method and was also used in implementations of the LS method to peptides.[9,75]

To estimate the extent of coverage of the reconstructed samples of the future chains one can generate a sample of the *entire* peptide (or loop) in the same way it is generated in the reconstruction process. Thus, starting from conformation i and discarding n_{init} conformations for equilibration, the sample can be of size n'_f (using g=10 fs) or of j consecutive samples of size n'_f, each starting from i with a different set of velocities. The $\Delta\alpha_k$ values (eq 13) of the dihedral angles (of both backbone and side chains) of this sample are then calculated and compared to the corresponding values obtained for the studied sample. The $\Delta\alpha_k$ values of the studied sample can also be compared with $\Delta\alpha_k$ values calculated during the reconstruction process itself for randomly chosen one or two conformations. These measures enable one to optimize the values of n_{init}, n'_f, and j. It should be pointed out, however, that in general we are interested in an entropy difference, $\Delta S^{\text{A}}_{m,n}$, between two microstates, where (as discussed later) the set n_{init}, n'_f, and j should be optimized simultaneously for both microstates; the result for $\Delta S^{\text{A}}_{m,n}$ is considered reliable if it is found to be stable for a large range of the parameters n_{init}, n'_f, and j. From now on we shall replace in most cases n'_f by the word unit.

II.7. Upper Bound and Exact Expressions for the Free Energy. In addition to $F^{\text{A}}(\rho^{\text{HS}}([\alpha_k]))$ (eq 17), which in practice is a lower bound, one can define an upper bound functional denoted F^{B}[47]

$$F^{\text{B}} = \frac{\int_m \rho^{\text{B}}[\rho^{\text{HS}} \exp[E/k_{\text{B}}T](E + k_{\text{B}}T \ln \rho^{\text{HS}})]d\alpha_1 \cdots d\alpha_K}{\int_m \rho^{\text{B}}[\rho^{\text{HS}} \exp[E/k_{\text{B}}T]]d\alpha_1 \cdots d\alpha_K} \tag{20}$$

Notice that (unlike F^{A}) the statistical reliability of estimating F^{B} decreases sharply with increasing system size. The inequalities $F^{\text{A}} \leq F \leq F^{\text{B}}$ will hold provided that the assumptions leading to eq 6 are valid. In this case F^{B} (like F^{A}) is increased by an additive constant (contributed by the bond stretching energy) which is cancelled in free energy differences of microstates. However, if deviations from these assumptions occur, F^{B} will be affected more significantly than F^{A} because $E/k_{\text{B}}T + \ln \rho^{\text{HS}}$ is exponiated in the numerator and denominator of eq 20; thus, to observe $F \leq F^{\text{B}}$ one might need to consider the bond-stretching entropy as well (see discussions in refs 46, 50, and 55).

As shown for fluids in ref 50, an *exact* expression for F, denoted F^{D}, is[55]

$$F^{\text{D}} = k_{\text{B}}T\ln\left(\frac{1}{Z_m}\right) = k_{\text{B}}T\ln[\int_m \rho^{\text{B}} \exp[F^{\text{HS}}/k_{\text{B}}T][d\alpha_k]] \tag{21}$$

where $[d\alpha_k] = d\alpha_1 \cdots d\alpha_K$ and $F^{\text{HS}}/k_{\text{B}}T = E([\alpha_k])/k_{\text{B}}T + \ln \rho^{\text{HS}}([\alpha_k])$. The above discussion for F^{B} also applies to F^{D}, where its estimation is statistically more reliable than that of F^{B} which is defined as a ratio of two summations similar to that defining F^{D}.

II.8. The Local States (LS) Method. With the LS method[44−46] (applied to an N-residue polyglycine with $6N = K$ backbone angles, α_k) the ranges $\Delta\alpha_k$ (eq 13) are divided into l equal segments, where l is the discretization parameter. These segments are denoted by ν_k (ν_k=1,l), where an angle α_k is represented by the segment ν_k to which it belongs, and a conformation i is expressed by the corresponding vector of segments [$\nu_1(i)$, $\nu_2(i)$, ...,ν_K (i)]. $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ can be estimated by $n(\nu_k, \cdots ,\nu_1)/\{n(\nu_{k-1}, \cdots ,\nu_1)[\Delta\alpha_k/l]\}$, where $n(\nu_k, \cdots ,\nu_1)$ is the number of times the *local state* [i.e., the vector $(\nu_k, \cdots ,\nu_1)$] appears in the sample. However, in practice, one uses smaller local states $(\nu_k,\nu_{k-1},...,\nu_{k-b})$ consisting of ν_k and its b preceding angles, where b is the correlation parameter. $n(\nu_k,\nu_{k-1},...,\nu_{k-b})$ lead to a set of transition probabilities $p(\nu_k|\nu_{k-1},..., \nu_{k-b})$ and *approximate* probability density, $\rho_i(b,l) = \prod_{k=1}^{K} p(\nu_k|\nu_{k-1},...,\nu_{k-b})/(\Delta\alpha_k/l)$, the larger are b and l the better the approximation (for enough statistics). The $\rho_i(b,l)$ lead to *rigorous* upper and lower bounds, S^{A} (eqs 16 and 18) and F^{A} (eq 17), respectively, where $\rho_i(b,l)$ replaces ρ^{HS}.

II.9. Calculation of Differences $S_m - S_n$. With QH, LS, and HSMC(D) calculation of $\Delta S_{mn} = S_m - S_n$ is based on the absolute values for each microstate. However, in section II.3 we have argued that the definition of a microstate m depends to a large extent on the simulation time t where *typically* m and its energy and entropy all grow with t. To be able to carry out a reliable estimation of ΔS_{mn} (ΔF_{mn}, etc.) we simulate both m and n for the same t looking for a range of t values where $\Delta F_{mn}(t)$, $\Delta S_{mn}(t)$, and $\Delta E_{mn}(t)$ are stable within the statistical errors [due to the simultaneous increase of $E_m(t)$, $E_n(t)$, etc.]. For the QH method such stable results constitute the best final answer. For the LS method, on the other hand, one can calculate $\Delta S^{\text{A}}_{mn}(b,l)$ [and $\Delta F^{\text{A}}_{mn}(b,l)$] for a set of improved approximations (by increasing b and l); if these differences converge within the statistical errors, the converged values are considered to be the correct differences due to cancellation of equal systematic errors in $S^{\text{A}}_m(b,l)$ and $S^{\text{A}}_n(b,l)$ (see discussion in section II.10). Notice that LS, unlike QH,[43] is applicable to a sample which covers several microstates and, in principle, even to a random coil.[49]

Obviously, if m is less stable than n, then the t values should be adjusted (i.e., decreased) to fit the stability of m. If m is significantly larger than n, then t_m should be large enough to allow adequate coverage of m $t_m \sim t_n[\Pi\Delta\alpha_k(m)/[\Pi\Delta\alpha_k(n)]$, where t_n is the time required to obtain an adequate sample for n. However, if $\Delta S_{mn}(t)$ increases monotonically, then it constitutes a lower bound. If the microstate is restrictive, e.g., side chains should populate a single rotamer, then the MD sample can be composed of several smaller samples each starts from the same structure with a different set of velocities. It should be pointed out that with LS and QH relatively large samples are required for obtaining converged TPs[46] and converged terms of the correlation matrix, σ (eq 5),[43] respectively. Therefore, one should verify that the samples remain in the original microstates and have not “escaped” to neighboring ones. We have developed methods for analyzing the stability of a microstate by calculating distribution profiles of dihedral angles.[9,75,77]

Unlike QH and LS, HSMC(D) is not based on gathering statistics from the studied sample; therefore, the required sample size is relatively small; also, F[HSMC(D)] (but not necessarily E and S[HSMC(D)]) can be obtained from a very small sample (even from a single conformation).[50] Therefore, in our studies the sample size for HSMC(D) is small, and it has been determined by the range of t values for which the average of E_m (E_n) is approximately constant. Again, one can envisage extreme cases where m is significantly larger than n, which would require increasing the sample size for m as described above for LS. With HSMC(D) the problem is to control the samples generated in the reconstruction process, as discussed in section II.6 and the next section (II.10). All these considerations are applicable to a peptide in different microstates (e.g., a helix, hairpin, or extended microstates[54,55]) as well as to a flexible surface loop, which populates significantly several microstates. In particular, the effect of sample size on $\Delta S_{mn} = S_m - S_n$ can be reduced, while controlling this effect with TI and the counting approaches is difficult (see discussion in ref 19).

II.10. Cancellation of Systematic Errors with HSMD. It should be pointed out that for any practical set of n_{init}, n'_f, j and bin sizes, $\delta\alpha_k$, the calculated S_m^{A} (S_n^{A}) will be approximate, and thus the corresponding difference, $S_m^{\text{A}} - S_n^{\text{A}}$, might be approximate as well. However, if $S_m^{\text{A}} - S_n^{\text{A}}$ is found to be stable for significantly improving approximations, the constant value can be considered to be the correct difference. Indeed, in the previous application of HSMD to peptides[55] and in the present study of a loop (see section III), relatively small values of n'_f and j have already led to stable differences, meaning that systematic errors in both S_m^{A} and S_n^{A} are comparable and thus are cancelled in $S_m^{\text{A}} - S_n^{\text{A}}$. This cancellation of relatively large systematic errors (discussed further below) makes HSMD an efficient procedure for peptides/loops.

To understand the basis for this cancellation, we examine first two one-dimensional harmonic microstates, i.e., two oscillators with equal mass and different spring constants f_1 and f_2. The *exact* entropy difference, $S_2 - S_1$, can be expressed in terms of the variances $<x_1^2>$ and $<x_2^2>$ of the corresponding coordinates

$$\Delta S_{2,1} = S_2 - S_1 = (1/2)k_{\text{B}} \ln\left(\frac{f_1}{f_2}\right) = k_{\text{B}}[\ln(<x_2^2>^{1/2}) - \ln(<x_1^2>^{1/2})] \quad (22)$$

One can estimate $\Delta S_{2,1}$ from two separate MD simulations, where the corresponding variances are calculated. If f_1 is significantly smaller than f_2 (i.e., f_1 defines a flatter parabola) and the same step size is used in both simulations a longer simulation will be required for f_1 than for f_2 to gain the same statistical precision. Therefore, if the same sample size is used for both microstates, then the statistical precision of $\Delta S_{2,1}$ will be determined mostly by that of S_1.

We now examine the entropy contributed by a backbone dihedral angle, α_k (denoted α for simplicity), in the course of the reconstruction process. α varies in microstates 1 and 2 within the ranges $\Delta\alpha_1$ and $\Delta\alpha_2$ (eq 13), which we denote Δ_1 and Δ_2, respectively. The crudest (but sometimes quite reliable) HSMD approximation for the corresponding difference in entropy, $\Delta S_0(\alpha)$, is

$$\Delta S_0(\alpha) = k_{\text{B}}[\ln\Delta_2 - \ln\Delta_1] \quad (23)$$

which is similar to that of eq 22 above (for brevity we shall omit α from the equations below). For better HSMD approximations, $\Delta S_0^{n_f}(l)$, we define the bins $\delta_1 = \Delta_1/l$ and $\delta_2 = \Delta_2/l$, where l is an increasing integer; the corresponding probabilities are $p_1^{n_f}(l)$ and $p_2^{n_f}(l)$ which are defined by n_{visit}/n_f (eq 14). One obtains

$$\Delta S_0^{n_f}(l) = k_{\text{B}}[\ln(p_1^{n_f}(l)/\delta_1) - \ln(p_2^{n_f}(l)/\delta_2)] = k_{\text{B}}\{\ln[p_1^{n_f}(l)/p_2^{n_f}(l)] + \ln(\Delta_2/\Delta_1)\}$$

or

$$\Delta S_0^{n_f}(l) = \Delta S^{n_f}(l) + \Delta S_0 \quad (24)$$

where $\Delta S^{n_f}(l)$ can be viewed as an anharmonic term. One can write $p_{\text{i}}^{\text{exact}}(l) = p_i^{n_f}(l)x_i^{n_f}(l)$ for $i = 1,2$, where $p_{\text{i}}^{\text{exact}}(l) = p_i^{n_f=\infty}(l)$ and $x_i^{n_f}(l)$ are thus factors (systematic errors) satisfying $x_i^{n_f}(l) \rightarrow 1$ for very large n_f; for a given l (bin) one obtains

$$\Delta S^{n_f}(l) = k_{\text{B}}\{\ln p_1^{\text{exact}}(l) - \ln p_2^{\text{exact}}(l) + \ln[x_2^{n_f}(l)/x_1^{n_f}(l)]\} \quad (25)$$

However, for large bins, δ (small l), one would expect to obtain probabilities that are close to the exact ones, $p_1^{\text{exact}}(l)$ and $p_2^{\text{exact}}(l)$ [i.e., $x_1^{n_f}(l)$ and $x_2^{n_f}(l)$ are $\sim$1] for a relatively small n_f due to adequate statistics, i.e., relatively large n_{visit} values. To obtain the exact probabilities (within the statistical errors) for decreased bin sizes, n_f should be increased adequately, which might increase computer time significantly. Thus, for practical values of n_f, $x_1^{n_f}(l)$ and $x_2^{n_f}(l)$ might differ significantly from 1 (i.e., large systematic errors). However, we argue that already for relatively small n_f, $x_2^{n_f}(l) \approx x_1^{n_f}(l)$, and the last logarithmic term (eq 25) becomes smaller than the *statistical* error leading to the correct value, $\Delta S(l)$, within the statistical error. To obtain the correct contribution (ΔS) of dihedral angle α to the *entropy difference* one has to define small enough bins, i.e., large enough $l_{\min}$, where for $l > l_{\min}$ $\Delta S(l)$ remains unchanged within the statistical error. As expected, $l_{\min}$ has to be smaller for a linear peptide than for a protein loop due to the restriction of loop closure, which requires relatively small bins (see sections III.1, 5, and 7).

The relation $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ stems from two reasons, where the first one is the fact that HSMD takes all interactions into account, and thus for a given n_f the future part of the chain is treated with the same level of approximation in both microstates. Second, because with MD the atoms are moved along their potential gradients, the simulations are equally efficient in both microstates. For peptides[55] the condition $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ occurs for much smaller n_f with HSMD than with HSMC[53] because the efficiency of the MC procedure used by us depends on the compactness of a structure (e.g., hairpin versus extended). Again, as for the parabolas above, if one microstate is significantly "flatter" than the other (i.e., larger $\Delta\alpha_k$ values), the required n_f value for obtaining convergence of ΔS will be determined mainly by the flatter microstate.

It should be noted that a $\Delta\alpha_k$ value of the studied sample might be significantly larger than the actual $\Delta\alpha_k$ available for α_k at step k of the reconstruction process of conformation i, due to geometrical constraints imposed by the constant "past", i.e., the partial structure reconstructed in the previous steps, 1....$k-1$. This limiting effect is expected to be more significant for dihedral angles than for bond angles; moreover, because at step k n_{visit} depends on mutual visits to the dihedral angle bin, $\delta\alpha_k$, and to its successive bond angle bin, $\delta\alpha_{k+1}$ (i.e., the modified eq 14 is $\rho^{\text{HS}}(\alpha_k,\alpha_{k+1}|\alpha_{k-1} \cdots \alpha_1) = n_{\text{visit}}/[n_f\, \delta\alpha_k \delta\alpha_{k+1}]$), $\delta\alpha_k$ and $\delta\alpha_{k+1}$ can be optimized to reduce S^{A} for a given n_f, which would lead to higher efficiency, i.e., to converged $S_m^{\text{A}} - S_n^{\text{A}}$ for smaller n_f (see sections III.4 and III.5). One can envisage a situation where for some side chains all rotamers are populated in one microstate, but only one rotamer is populated in the other microstate and vice versa (see section II.7). These differences might compensate each other in $S_m^{\text{A}} - S_n^{\text{A}}$; therefore, evaluating the reconstruction calculations should be carried out with extra caution. Again, the ultimate test for accuracy is the occurrence of stable $S_m^{\text{A}} - S_n^{\text{A}}$ values for increasing n_f and decreasing bin sizes, as previously discussed.

As mentioned above, with the MC method used by us[53] an open peptide structure (e.g., the extended microstate of a peptide) is simulated more efficiently than a compact hairpin structure and therefore relatively large n_f was needed to achieve $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ within the statistical errors. Thus far we have studied by HSMD microstates of three systems, degaglycine, $NH_2(Val)_2(Gly)_6(Val)_2CONH_2$,[55] and in this paper the 7-residue loop of α-amylase in vacuum and implicit solvent. In all these studies the cancellation of systematic errors has been found to occur for relatively small n_f, which has been verified by comparing entropy differences obtained for a wide range of n_f values. For decaglycine, for example, n_f ranges between 500 and 24 000, where $n_f = 500$ (5 ps) leads to the correct results, and HSMD is thus ~100 times more efficient (in terms of computer time) than HSMC.[53] We expect this cancellation of errors to occur also for models of peptides and loops in explicit water.

III. Results and Discussion

III.1. Simulation Details for the Loop in Vacuum. An MD simulation (at 300 K) starting from the free PDB structure has led to a stable sample of size 600 (where a structure is added to the sample every 0.5 ps) around the PDB structure with an average energy of ~ -138 kcal/mol. However, in simulations of this size, starting from the bound PDB structure, the initial energy (~ -98 kcal/mol) was decreased constantly and significantly; we have found this energy to stabilize around -110 kcal/mol only after a very long MD run. Because we are interested in studying the stability of the bound microstate in the free protein (see sections I.7 and II.1), its sample was generated by combining partial samples obtained from short MD runs each started from the PDB bound structure with a different set of velocities. The average energy of these short samples remained close to -98 kcal/mol. This procedure can be used with HSMC(D), which operates on small samples (and in an extreme case even on a single structure), while it is less effective with the LS[9] and the QH methods, which require much larger samples, as discussed earlier (section II.9). Generating such a combined sample will be useful also for studying the entropy (and energy) of a transition state.

The free and bound samples and the reconstruction simulations (future samples) were carried out with the velocity-Verlet algorithm[28] based on a time step of 1 fs, where the Berendsen[28] heat bath controlled the temperature. Cut-offs on long-range interactions were not imposed, and in the reconstruction process a structure was added to the sample every $g = 10$ fs, where the $n_{\text{init}} = 250$ initial structures (2.5 ps) were discarded for equilibration. The future samples were generated for four bin sizes, $\delta = \Delta\alpha_k/30$, $\Delta\alpha_k/15$, $\Delta\alpha_k/10$, and $\Delta\alpha_k/5$, centered at α_k (i.e., $\alpha_k \pm \delta/2$) (eqs 13 and 14). If the counts of the smallest bin are smaller than 50, then the bin size is increased to the next size and if necessary to the next one, etc. In the case of zero counts, n_{visit} is taken to be 1; however, zero counts is a very rare event. In this context it should be pointed out that if one had tried to build loop structure i by selecting angles at random within the ranges $\alpha_k \pm \delta/2$, the constructed structure would differ from i and in the case of a loop would not satisfy the loop closure condition leading to a very high bond stretching energy. Therefore, the smallest bin chosen for a loop ($\Delta\alpha_k/30$) is smaller than that used for the linear peptides[55] ($\Delta\alpha_k/15$). Notice, however, that this structural deviation from i would affect both microstates, and the bins used are the largest that still lead to converging results of ΔS^{A}.

For each microstate, two sets of results were calculated, one is based on unit $n'_f = 250$ (2.5 ps) and n_f values of 250 (j=1), 500 (j=2), 750 (j=3), and 1250 (j=5). The second set is based on unit $n'_f = 1000$ (10 ps) and n_f values of 1000 (j=1), 2000 (j=2), 4000 (j=4), and 8000 (j=8) (see section II.6). These sets that lead to an increasing coverage of the studied microstates, enable one to examine the convergence of $S^{\text{A}}(n'_f, j)$ as well as $\Delta S_{mn}^{\text{A}}(n'_f, j)$, which is our main interest.

Table 1. Differences $\Delta\alpha_k$ (in deg) between the Minimum and Maximum Values of Dihedral Angles in the Free and Bound Samples in Vacuum[a]

	free loop				bound loop			
	studied sample		1 × 2.5 ps (5 × 2.5 ps)		studied sample		1 × 2.5 ps (5 × 2.5 ps)	
residue	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$
Gly 1	46	74	43 (52)	61 (108)	43	90	51 (48)	96 (74)
His 2	75	88	62 (98)	71 (189)	89	61	78 (87)	51 (68)
Gly 3	58	112	64 (298)	80 (109)	68	72	54 (73)	63 (88)
Ala 4	99	89	73 (103)	70 (87)	91	94	71 (73)	53 (76)
Gly 5	99	105	77 (83)	80 (139)	84	101	58 (70)	43 (60)
Gly 6	112	85	88 (118)	85 (67)	59	64	43 (57)	42 (46)
Ser 7	69	54	52 (88)	65 (65)	81	44	42 (58)	35 (39)
χ^1 (His)	58		36 (66)		45		44 (50)	
χ^2(His)	145		117 (116)		103		62 (96)	
χ^1 (Ser)	155		55 (163)		39		31 (51)	

[a] $\Delta\alpha_k$ are defined in eq 13. The studied samples of $n = 600$ conformations were generated with the AMBER force field by retaining a conformation every 500 fs. The 1 × 2.5 ps samples (of 250 conformations each) were started from two chosen conformations of the free and bound (studied) samples, by retaining a conformation every 10 fs and ignoring the first 250 conformations for equilibration. The sample denoted (5 × 2.5 ps) consists of five 2.5 ps samples (altogether 1250 conformations) each started from the chosen structure with a different set of velocities where the initial 250 conformations are ignored for equilibration.

III.2. Entropy Results in Vacuum. In Table 1 we present the values of $\Delta\alpha_k$ (eq 13) for the free and bound microstates obtained from the corresponding MD samples. These values suggest that the two samples indeed are concentrated in conformational space. Table 2 contains two sets of results for the entropy, TS^A (eq16) based on units of 2.5 and 10 ps for the free and bound microstates. As mentioned in section III.1, these results were calculated for four different future sample sizes, n_f and four bin sizes; however, the extent of convergence is demonstrated by the results obtained for the three smallest bin sizes, $\Delta\alpha_k/10$, $\Delta\alpha_k/15$, and $\Delta\alpha_k/30$, which are presented in the table. The statistical errors were obtained from the fluctuations and results obtained for partial samples. These results were obtained without considering the Jacobian ($\Pi_j \sin(\theta_j)$) (see discussion following eq 6), which enables one to compare them to those obtained previously for peptides[55] without considering the Jacobian as well. As shown later in section III.4, the contribution of the Jacobian to the entropy cancels out to a very good approximation in entropy and free energy differences—our main interest. Therefore, ignoring the Jacobian, which increases the statistical errors (hence requires larger samples), is justified.

One would expect S^A to decrease with decreasing the bin and increasing n_f—an expectation, which is fully satisfied by the results of Table 2. In particular, for a given n_f, S^A always decreases as the bin is decreased; however, a complete convergence occurs only for the free loop (unit=2.5 ps), where $TS^A(\Delta\alpha_k/15,n_f=1250)$ is equal to $TS^A(\Delta\alpha_k/30,n_f=1250)$ within the relatively large statistical errors; for the other cases the deviation from convergence are small. Convergence of the two best results (i.e., for the largest n_f) for each bin occurs only for unit = 10 (for both microstates), while for unit = 2.5 ps the deviations from full convergence are again small $T[S^A(\delta,n_f=750) - S^A(\delta,n_f=1250) \leq 1.2$ kcal/mol for all δ]. It should be pointed out that the errors for the bound loop are smaller than those for the free loop probably due to the fact that the sample of the bound loop consists of several subsamples that were generated from the same initial structure.

The HSMD results for the entropy are compared in the table to those obtained with the LS and QH methods from larger MD samples of 5000, 8000, and 10 000 conformations. These samples consist of several subsamples each started from the same structure with a different set of velocities, where a conformation was retained every 50 fs. The QH results for TS exceed the HSMD values by 12.8 and 11.3 kcal/mol for the free and bound microstate, respectively, which is in accord with S^{QH} being an upper bound; these differences are probably also affected (i.e., increased) by the significantly larger samples used for QH than for HSMD (see discussion in section II.9). The LS results (calculated for b=1, l=10), which also constitute upper bounds, are larger than the corresponding QH values, as was also found in previous studies.[53−55]

III.3. Free Energy Results in Vacuum. Results for the free energy functional, F^A (eq 17), its fluctuation, σ_A (eq 19), and the energies are presented in Table 3. These results are given only for the smallest bin, $\Delta\alpha_k/30$ and unit = 10, because F^A values for the other bins can be obtained from the entropies of Table 2 and the energies provided in the bottom of Table 3. F^A (like S^A) does not change (within the errors bars) as n_f is increased from 5000 to 8000, and the slight (expected) decrease of the central values of σ_A with increasing n_f is, however, insignificant within the error bars. As expected, the QH and LS results for F underestimate the correct values, and the central values of the energy fluctuations are always larger than those for $\sigma_A(n_f=8000)$. Finally, the table shows the differences in free energy and energy between the free and bound microstates. It is evident that the free energy differences, ΔF^A, are all equal within the statistical errors, and they are also equal to the energy difference, ΔE and ΔF^A_{LS}, obtained with LS. This suggests that the ΔS^A results are ~0, as indeed shown in the next section, III.4, i.e., the free microstate is more stable than the bound one by ~38.8 kcal/mol which is contributed mainly by ΔE.

The results for F^B (eq 20) are not provided in the table because they do not behave as expected, i.e., they do not decrease as n_f is increased and the bin is decreased. This "misbehavior" can be attributed to a too small sample size n and to the fact that the bond stretching energy is included in the potential energy, while the corresponding entropy is not taken into account in ρ^{HS} (eq 15) (see discussion in ref 55). Still, the results obtained for F^B are always larger than those of F^A and thus probably provide upper bounds; the deviations, however, are relatively large ($F^B = -191.7$ and -150.2 kcal/mol, deviating from F^A by ~12 and 14 kcal/mol for the free and bound microstates, respectively). Due to the almost convergence of the F^A values it is plausible to assume that the F^B results do not lead to improved approximations for the free energy, i.e., the average values, $F^M = (F^A+F^B)/2$ are probably less reliable than those of F^A. Notice, however, that $\Delta F^B = -39.7$ is only 1 kcal/mol

Table 2. HSMD Results (in kcal/mol) for the Entropy, TS^A (Eqs 16 and 18), at $T = 300$ K Calculated from Samples of 600 Conformations of the Free and Bound Microstates in Vacuum[a]

	free loop				bound loop			
	unit = 2. 5 ps (250)		unit = 10 ps (1000)		unit = 2.5 ps (250)		unit = 10 ps (1000)	
Δ	$n_f(j)$	TS^A	$n_f(j)$	TS^A	$n_f(j)$	TS^A	$n_f(j)$	TS^A
$\Delta\alpha_k/10$	250 (1)	72.1 (3)	1000 (1)	68.6 (2)	250 (1)	71.01 (3)	1000 (1)	67.78 (3)
$\Delta\alpha_k/10$	500 (2)	69.5 (3)	2000 (2)	68.2 (2)	500 (2)	68.84 (3)	2000 (2)	67.35 (3)
$\Delta\alpha_k/10$	750 (3)	68.6 (3)	4000 (5)	68.1 (1)	750 (3)	67.81 (4)	4000 (5)	67.28 (3)
$\Delta\alpha_k/10$	1250 (5)	**67.9 (2)**	8000 (8)	**68.0 (1)**	1250 (5)	**67.15 (3)**	8000 (8)	**67.26 (2)**
$\Delta\alpha_k/15$	250 (1)	71.9 (3)	1000 (1)	67.4 (2)	250 (1)	70.90 (3)	1000 (1)	66.89 (3)
$\Delta\alpha_k/15$	500 (2)	69.0 (3)	2000 (2)	66.8 (2)	500 (2)	68.46 (5)	2000 (2)	66.28 (3)
$\Delta\alpha_k/15$	750 (3)	67.6 (3)	4000 (5)	66.7 (1)	750 (3)	67.14 (4)	4000 (5)	66.24 (3)
$\Delta\alpha_k/15$	1250 (5)	**66.6 (2)**	8000 (8)	**66.7 (1)**	1250 (5)	**66.10 (3)**	8000 (8)	**66.22 (3)**
$\Delta\alpha_k/30$	250 (1)	71.9 (2)	1000 (1)	67.2 (2)	250 (1)	70.89 (3)	1000 (1)	66.77 (4)
$\Delta\alpha_k/30$	500 (2)	68.9 (2)	2000 (2)	66.3 (2)	500 (2)	68.42 (4)	2000 (2)	65.97 (4)
$\Delta\alpha_k/30$	750 (3)	67.5 (2)	4000 (5)	65.9 (1)	750 (3)	67.06 (4)	4000 (5)	65.60 (4)
$\Delta\alpha_k/30$	1250 (5)	**66.3 (2)**	8000 (8)	**65.8 (1)**	1250 (5)	**65.91 (3)**	8000 (8)	**65.51 (4)**
TS^{QH}		78.61 (7)		78.61 (7)		76.8 (2)		76.8 (2)
TS^{LS}		91.6 (4)		91.6 (4)		90.9 (7)		90.9 (7)

[a] The bin sizes are $\delta = \Delta\alpha_k/l$ (eq 13). The two units of 2.5 and 10 ps used are also defined (in parentheses) by their number of conformations, 250 and 1000, respectively. n_f denotes the sample size of the future chains used in the reconstruction process, n_f = unit × j, where j is the number of simulations of unit size applied at each reconstruction step. The statistical errors are given in parentheses, e.g., 66.3 (2) = 66.3 ± 0.2. S^{QH} is the quasi-harmonic entropy (eq 5), and S^{LS} (eqs 16 and 18 and section II.8) is S^A obtained by the local states method using $b = 1$ and discretization parameter $l = 10$; these results were obtained from larger samples (for details see text). All calculations were carried out with the AMBER force field. The entropy is defined up to an additive constant that is the same for both microstates.

Table 3. HSMD Results at $T = 300$ K for the Free Energy, F^A, the Interaction Energy, E_{int}, Their Fluctuations, and ΔF^A and ΔE for the Free and Bound Microstates in Vacuum[a]

	free loop		bound loop		free−bound	
n_f	$-F^A$	σ_A, σ_E	$-F^A$	σ_A, σ_E	ΔF^A	ΔE_{int}
1000	204.7 (3)	4.4 (3)	165.79 (8)	4.5 (3)		−38.9 (4)
2000	203.7 (3)	4.3 (3)	165.0 (1)	4.4 (2)		−38.7 (4)
4000	203.3 (3)	4.3 (3)	164.62 (6)	4.4 (2)		−38.7 (3)
8000	203.3 (2)	4.2 (3)	164.54 (6)	4.4 (2)		−38.7 (3)
$-F^{QH}$	216.1 (1)		175.8 (3)			−40.2 (4)
$-F^{LS}$	229.1 (4)		190.0 (7)			−39.1 (5)
$-E_{int}$	137.5 (3)	4.4 (3)	99.02 (5)	4.49 (4)		−38.4 (3)

[a] F^A (eq 17) is a lower bound of the free energy, and σ_A (eq 19) is its fluctuation. The results were obtained from samples of $n = 600$ conformations for the smallest bin size, $\delta = \Delta\alpha_k/30$, unit = 10 ps, and all future sample sizes n_f. F^{QH} (see eq 5) and F^{LS} (eq 17 and section II.8) are free energies obtained by the quasi-harmonic approximation and the local states method (b=1, l=10), respectively, and are based on larger samples (see text). The average potential energy, E_{int}, of the studied samples appears in the bottom row; σ_E is the energy fluctuation (these results are in kcal/mol). All free energies are in kcal/mol and are defined up to the same additive constant for both microstates. All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

Table 4. Entropy Differences, $T\Delta S^A = T[S^A_{free} - S^A_{bound}]$ (in kcal/mol) at $T = 300$ K in Vacuum[a]

	unit = 2.5 ps (250)			unit = 10 ps (1000)		
	n_f	$T\Delta S^A$	$T\Delta S^A$ (Jacobian)	n_f	$T\Delta S^A$	$T\Delta S^A$ (Jacobian)
$\Delta\alpha_k/15$	250	1.0 (2)	1.2 (1)	1000	0.5 (2)	0.6(2)
$\Delta\alpha_k/15$	500	0.5 (2)	0.6 (2)	2000	0.5 (2)	0.7 (1)
$\Delta\alpha_k/15$	750	0.5 (2)	0.6 (2)	4000	0.5 (1)	0.6 (1)
$\Delta\alpha_k/15$	1250	**0.5 (1)**	**0.6 (1)**	8000	**0.5 (1)**	**0.6 (1)**
$\Delta\alpha_k/30$	250	1.0 (2)	1.1 (1)	1000	0.5 (2)	0.6 (1)
$\Delta\alpha_k/30$	500	0.5 (2)	0.6 (2)	2000	0.3 (2)	0.4 (1)
$\Delta\alpha_k/30$	750	0.4 (2)	0.5 (2)	4000	0.3 (1)	0.4 (1)
$\Delta\alpha_k/30$	1250	**0.4 (1)**	**0.5 (1)**	8000	**0.3 (1)**	**0.4 (1)**
$T\Delta S^{QH}$		1.8 (2)			1.8 (2)	
$T\Delta S^{LS}$		0.6 (3)			0.6 (3)	

[a] S^A is an upper bound of the entropy (eqs 16 and 18). The results for $T\Delta S^A$ were obtained from samples of $n = 600$ conformations for the two smallest bins, $\delta = \Delta\alpha_k/15$ and $\delta = \Delta\alpha_k/30$, unit = 2.5 and 10 ps, and all the future sample sizes, n_f. We also present $T\Delta S^A$ results obtained by HSMD, where S^A is calculated with the Jacobian (see discussion following eq 6). $T\Delta S^{QH}$ (eq 5) and $T\Delta S^{LS}$(eqs 16 and 18 and section II.8) are entropy differences calculated by the quasi-harmonic approximation and the local states method (b=1, l=10); they are based on larger samples (see text). All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

smaller than ΔF^A. We also calculated the corresponding F^D values, −196.9 and −154.0 kcal/mol (eq 21), which are smaller than the related F^B results but are larger than those for F^M, leading to $\Delta F^D = -42.9$ kcal/mol. While it would be desirable to have converging results for F^B and F^D, we demonstrate below that reliable *differences* in S (and F) can be obtained from differences in S^A (and F^A).

III.4. Entropy Differences in Vacuum. Because computer time increases linearly with n_f, it is of interest to check the effect of decreased n_f on entropy differences. In Table 4 results are presented for $T\Delta S^A = T[S^A_{free} - S^A_{bound}]$ calculated for the two smallest bins, the four n_f values, and unit = 2.5 and 10 ps. We also present results for $T\Delta S^A$ calculated with the Jacobian. The table reveals that the corresponding results obtained with and without the Jacobian are equal within the error bars. This is important because the calculations without the Jacobian have converged statistically already for samples of 400 conformations, while including the Jacobian required increasing the sample size to 600. The results for unit = 2.5 and $n_f \geq 500$ are converged (within the error bars) with respect to both n_f and the two bin sizes. The fact that the

$T\Delta S^{A}$ value obtained for unit = 2.5 ps and n_f = 500 is equal to that obtained for a four times larger unit (of 10 ps) and for a 16 times larger n_f (8000) suggests that $T\Delta S^{A} = 0.3 \pm 0.1$ kcal/mol is the correct result.

This stems from the cancellation (in $T\Delta S^{A}$) of approximately equal systematic errors for both microstates, as discussed in section II.10. Thus, Table 2 shows that the worst approximations that still lead to the correct $T\Delta S^{A}$ differ from the best ones by $TS^{A}(\Delta\alpha_k/15,n_f{=}500) - TS^{A}(\Delta\alpha_k/30,n_f{=}8000)$ = 3.2 and 3.0 kcal/mol for the free and bound microstates, respectively; these differences constitute lower bounds because the correct TS values might be significantly smaller than $TS^{A}(\Delta\alpha_k/30,n_f{=}8000)$. The table shows that the difference obtained by the LS method [0.6 (3) kcal/mol] is equal to that obtained by HSMD, while QH leads to a significantly higher difference, 1.8 (2) kcal/mol. However, this good LS result might be accidental as unreliable differences were obtained by LS for the extended, helix, and hairpin microstates of decaglycine.[53,55]

The similar results obtained for unit = 2.5 and 10 ps suggest that already for unit =2.5 (n_f=500) the coverage of both microstates by the future chains is adequate and that for unit = 10 and n_f = 8000 the future chains still remain within these microstates. To get an idea about the extent of this coverage, we selected a structure from each of the two studied samples from which MD simulations of the *entire* loop were started (see the last paragraph of section II.6). Two samples of 250 conformations and two samples of 5 × 250 = 1250 conformations were generated in the same way the future chains are simulated during the reconstruction process, i.e., 1250 consists of five subsamples (i.e., units) of 250 conformations, each starting from the initial structure with a different set of velocities where the first 250 structures are ignored for equilibration. In these simulations a structure is retained (as in the reconstruction process) every g = 10 fs (unlike the two studied samples that were generated with g=500 fs). The $\Delta\alpha_k$ results (eq 13) for the dihedral angles for these samples are presented in Table 1 which shows that in most cases the results for 2.5 ps are slightly smaller than the corresponding results obtained for the studied samples, while the (expected) larger results for 5 × 250 are still close to those of the studied samples; this applies also to the side chains. Deviations from this picture occur for 5 × 250, where $\Delta\psi$ and $\Delta\varphi$ of His^2 and Gly^3, respectively, are significantly larger than the corresponding values of the studied samples. This picture suggests that the reconstruction simulations cover adequately the two studied microstates.

In view of the discussion in section II.10 we have also tried to optimize the bins' sizes. As pointed out there, the values of $\Delta\alpha_k$(dihedral) in Table 1 (for the entire samples) are expected to overestimate the actual $\Delta\alpha_k$ available for α_k at step k of the reconstruction process. Therefore, a relatively large number of visits of α_k(dihedral) to its bin $\delta\alpha_k = \Delta\alpha_k/l$ are not followed by visits of α_{k+1}(bond angle) to its bin, α_{k+1}/l; thus, the number of counts (at both $\delta\alpha_k$ and $\delta\alpha_{k+1}$) is relatively small, while $\Delta\alpha_k$(dihedral)/l is large, leading to small $\rho^{\mathrm{HS}}(\alpha_k,\alpha_{k+1}|\alpha_{k-1} \cdots \alpha_1) = n_{\mathrm{visit}}/[n_f\delta\alpha_k\delta\alpha_{k+1}]$, i.e., to a large contribution, $-k_{\mathrm{B}}\ln\rho^{\mathrm{HS}}$, to the entropy. This undesirable effect can be reduced by decreasing the values of $\Delta\alpha_k$(dihedral) used for defining the dihedral bins or increasing the values of $\Delta\alpha_k$(bond angle) used for defining the bond angles' bins. We have adopted the latter option by increasing *all* of the bond angles bins to 50°/l (typically $\Delta\alpha_k$(bond angle) ranges from 20 to 25°) and applied HSMD with a relatively small unit = 1 ps and small n_f = 100, 200, 300, and 400. The results for $T\Delta S^{A}$ appear in Table 5 and are shown to be very close to those of Table 4, which suggests that HSMD can be optimized further leading to a further reduction in computer time.

Table 5. Entropy Differences, $T\Delta S^{A} = T[S^{A}_{\mathrm{free}} - S^{A}_{\mathrm{bound}}]$ (in kcal/mol) at T = 300 K in Vacuum Using Equal Bins for the Bond Angles[a]

	unit = 1 ps (100)	
	n_f	$T\Delta S$
$\Delta\alpha_k$/10	100	0.7 (1)
$\Delta\alpha_k$/10	200	0.7 (1)
$\Delta\alpha_k$/10	300	0.7 (2)
$\Delta\alpha_k$/10	400	**0.5 (1)**
$\Delta\alpha_k$/15	100	0.6 (1)
$\Delta\alpha_k$/15	200	0.6 (1)
$\Delta\alpha_k$/15	300	0.6 (2)
$\Delta\alpha_k$/15	400	**0.5 (2)**
$\Delta\alpha_k$/30	100	0.6 (1)
$\Delta\alpha_k$/30	200	0.6 (1)
$\Delta\alpha_k$/30	300	0.6 (2)
$\Delta\alpha_k$/30	400	**0.4 (2)**

[a] S^{A} is an upper bound of the entropy (eqs 16 and 18). The results for $T\Delta S^{A}$ were obtained from samples of n = 600 conformations for the three smallest bins, using unit = 1 ps. The bond angles bins are δ = 50°/l, while for the dihedral angles they are $\delta = \Delta\alpha_k/l$ (eq 13), where l = 10, 15, and 30. n_f is the sample size of the future chains in the reconstruction procedure. All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

These results support the conclusions obtained for peptides[55] that correct differences ΔS^{A}_{mn} can be obtained for relatively short reconstruction simulations, which leads to considerable savings in computer time. In fact, reconstruction of a structure based on n_f = 500 and 100 requires, respectively, ~30 and 14 min CPU on a 2.1 GHz Athlon processor. This time can be reduced by a factor of 2 if the MD integration is carried out with a time step of 2 fs (rather than 1 fs). Due to strong correlations among the dihedrals and bond angles within a microstate, it might be possible to treat four successive angles (two dihedrals and two bond angles) rather than two angles considered presently at each reconstruction step. One can increase efficiency further by applying a cutoff on long-range interactions and running the simulations on the best machines available to date. One would seek to decrease computer time further by considering the conformational restraints imposed by the loop closure condition on the pair of dihedral angles, α_k, and its successive bond angle, α_{k+1}, at each reconstruction step. However, in spite of this restraint the fluctuations in these angles (partially due to bond stretching) are significant in all reconstruction steps besides the last two (K- 4, K-3 and K-2, K-1, where K is the last angle in the loop). While one could probably ignore the reconstruction of these two last steps in both microstates (as long as differences, ΔS^{A} are of interest), the gain in

Table 6. Differences $\Delta\alpha_k$, (in deg) between the Minimum and Maximum Values of Dihedral Angles in the Free and Bound Samples in Solvent[a]

	free loop (solvent)				bound loop (solvent)			
	entire sample		1 × 5 ps (10 × 5 ps)		entire sample		1 × 5 ps (10 × 5 ps)	
residue	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$
Gly 1	76	153	54 (101)	122 (175)	92	148	85 (109)	125 (200)
His 2	139	130	114 (140)	95 (360)	125	105	105 (122)	101 (162)
Gly 3	175	124	88 (285)	99 (176)	80	95	100 (202)	139 (151)
Ala 4	131	94	68 (170)	75 (89)	143	288	134 (168)	125 (360)
Gly 5	107	100	89 (119)	111 (131)	199	360	130 (226)	122 (360)
Gly 6	126	109	154 (351)	97 (134)	285	267	205 (293)	357 (357)
Ser 7	83	64	75 (87)	56 (71)	243	109	127 (188)	90 (114)
X^1 (His)	55		43 (66)		53		33 (77)	
X^2 (His)	130		102 (161)		108		95 (130)	
X^1 (Ser)	317		60 (188)		321		51 (167)	

[a] $\Delta\alpha_k$ are defined in eq 13. The studied samples of $n = 500$ conformations were generated by retaining a conformation every 500 fs. The 1 × 5 ps samples (of 500 conformations each) were started from two chosen conformations of the free and bound (studied) samples, by retaining a conformation every 10 fs and ignoring the first 250 conformations for equilibration. The sample denoted (10 × 5 ps) consists of ten 5 ps samples (altogether 5000 conformations) each started from the chosen structure with a different set of velocities where the initial 250 conformations are ignored for equilibration. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA.

Table 7. HSMD Results (in kcal/mol) for the Entropy, TS^{A} (Eqs 16 and 18) at $T = 300$ K Calculated from a Sample of 200 Conformations of the Free and Bound Microstates in Solvent[a]

bin size	n_f	$TS^{\mathrm{A}}_{\mathrm{free}}$	$TS^{\mathrm{A}}_{\mathrm{bound}}$	$T[S^{\mathrm{A}}_{\mathrm{free}} - S^{\mathrm{A}}_{\mathrm{bound}}]$
$\Delta\alpha_k/15$	625	70.7 (1)	71.0 (2)	- 0.3 (1)
$\Delta\alpha_k/15$	1250	69.9(1)	70.3 (2)	−0.4 (1)
$\Delta\alpha_k/15$	2500	69.7 (2)	70.1 (2)	−0.4 (1)
$\Delta\alpha_k/15$	5000	**69.6 (2)**	**70.0 (2)**	−**0.4 (1)**
$\Delta\alpha_k/30$	625	70.6 (1)	70.9 (2)	−0.3 (1)
$\Delta\alpha_k/30$	1250	69.4 (1)	69.8 (2)	−0.4 (1)
$\Delta\alpha_k/30$	2500	68.8 (2)	69.2 (2)	−0.4 (1)
$\Delta\alpha_k/30$	5000	**68.4 (2)**	**68.8 (2)**	−**0.4 (1)**
$\Delta\alpha_k/45$	625	70.6 (1)	70.9 (2)	−0.3 (1)
$\Delta\alpha_k/45$	1250	69.4 (2)	69.7 (2)	−0.4 (1)
$\Delta\alpha_k/45$	2500	68.7 (2)	69.1 (2)	−0.4 (1)
$\Delta\alpha_k/45$	5000	**68.2 (2)**	**68.5 (1)**	−**0.4 (1)**
TS^{QH}		89 (1)	92 (1)	−3 (1)
TS^{LS}		100 (1)	108 (1)	−8 (1)

[a] The bin sizes for the bond angles are $\delta = 100°/l$ (i.e., $\Delta\alpha_k$=100°) and are $\delta = \Delta\alpha_k/l$ for the other angles, where $\Delta\alpha_k$ is defined in eq 13. n_f, the sample size of the future chains generated in the reconstruction process, is based on unit = 500 conformations (5 ps). S^{QH} is the quasi-harmonic entropy (eq 5), and S^{LS} (eqs 16 and 18 and section II.8) is S^{A} obtained by the local states (LS) method using $b = 1$ and the discretization parameter, $l = 10$; these results were obtained from larger samples (for details see text). The entropy is defined up to an additive constant which is the same for both microstates. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA. The statistical error is defined in footnote *a* of Table 2.

computer time would be small. Finally, while the structure of TINKER makes it a very convenient tool for developing new programs, the code is not very efficient, decreasing the performance of HSMD as well.

III.5. Results for the Loop in Implicit Water. These MD simulations are based on the AMBER force field[68] and the GB/SA solvation model of Still and co-workers[69] which (like AMBER) is implemented within TINKER.[70] The samples for the free and bound microstates were generated at $T = 300$ K in a similar way to those in vacuum with some changes: the sample size is $n = 500$ (rather than 600), and the MD step size was increased to 2 fs, where bonds involving hydrogens were frozen to their ideal values by using the RATTLE algorithm;[28] also, the smallest bin size in the reconstruction process was decreased to $\delta = \Delta\alpha_k/45$, and thus the other three bins are $\Delta\alpha_k/30$, $\Delta\alpha_k/15$, and $\Delta\alpha_k/10$.

The $\Delta\alpha_k$ results for the free and bound samples appear in Table 6, and they are shown to be larger than the corresponding values obtained for the loop in vacuum in Table 1. This increase in $\Delta\alpha_k$ is expected due to the protein−solvent interactions that lead to an increase in the loop flexibility, hence to its larger entropy. The table also reveals that in most cases the $\Delta\alpha_k$ values of the bound microstate are larger than their counterparts in the free microstate and in some cases, for $\Delta\varphi$ of Gly[5], Gly[6], and Ser[7] and $\Delta\psi$ of Ala[4], Gly[5], and Gly[6], the difference is significant where $\Delta\alpha_k$(bound) ranges from 200 to 360°. This might lead to the conclusion that the entropy of the bound microstate is significantly larger than that of the free microstate. However, one should bear in mind that the α_k are highly correlated, and in the case of small residues, such as Gly and Ala significant simultaneous changes in neighbor dihedral angles can lead mainly to small *localized* conformational changes and thus to a relatively narrow "pipe" of low entropy (see section II.3). One has also to verify that the large $\Delta\alpha_k$ values of the bound microstate do not lead to an overlap of the two microstates. Comparing structures of the two samples generated at the same time along the trajectories shows that the energies differ by ~25 kcal/mol (see also Table 8), the rmsd of all heavy atoms is ~2.2 Å, and the corresponding dihedral angles are different and in some cases significantly different (e.g., 109 vs −45° for ψ of Ala[4]).

Using the (relatively large) $\Delta\alpha_k$ values of Table 6 as a basis for defining the bin sizes for the reconstruction process will lead to a set of results $S^{\mathrm{A}}_{\mathrm{bound}}$ with a lower level of approximation than the corresponding results of $S^{\mathrm{A}}_{\mathrm{free}}$; consequently, relatively small bins and large n_f values will be needed to obtain a set of converging results for the difference $\Delta S^{\mathrm{A}} = S^{\mathrm{A}}_{\mathrm{free}}(\Delta\alpha_k/l,n_f) - S^{\mathrm{A}}_{\mathrm{bound}}(\Delta\alpha_k/l,n_f)$. Indeed, preliminary calculations have led to decreasing nonconverging results where $T\Delta S^{\mathrm{A}} \sim -1$ kcal/mol for the best approximation, $n_f = 5000$ and $\delta = \Delta\alpha_k/45$. To obtain sets of results of $S^{\mathrm{A}}_{\mathrm{bound}}$ which are on the same level of approximation as those of $S^{\mathrm{A}}_{\mathrm{free}}$ we have defined (similar to the vacuum case in section III.4) a uniform set of bins for the bond angles (for both microstates) as $100°/l$, where $l = 45$, 30, 15, and 10, while the dihedral angles' bins, $\Delta\alpha_k/l$, are based on the $\Delta\alpha_k$ values of Table 6.

III.6. Entropy in Implicit Water. The computations with GB/SA are much more time-consuming than those carried out in vacuum; therefore, we performed only one set of calculations based on unit = 500 (5 ps) with $n_f = 625$, 1250,

Table 8. HSMD Results at $T = 300$ K for the Free Energy, F^A, the Potential Energy, E_{int}, Their Fluctuations, and the Differences, ΔF^A and ΔE_{int}, between the Free and Bound Microstates in Solvent[a]

	free loop		bound loop		free−bound	
n_f	$-F^A$	$\sigma_{A,}\ \sigma_E$	$-F^A$	$\sigma_{A,}\ \sigma_E$	ΔF^A	ΔE
625	939.0 (1)	3.7 (1)	913.4 (2)	3.7 (2)		−25.6 (1)
1250	937.8 (2)	3.7 (1)	912.3 (2)	3.7 (1)		−25.5 (1)
2500	937.1 (2)	3.8 (1)	911.6 (1)	3.7 (2)		−25.5 (1)
5000	936.6 (2)	3.8 (1)	911.1 (2)	3.7 (2)		−25.5 (1)
$-F^{QH}$	955 (1)		935 (1)			−21 (1)
$-F^{LS}$	966 (2)		950 (2)			−16 (3)
$-E_{int}$	868.3 (5)	4.1 (1)	842.6 (3)	4.0 (1)		−25.7 (1)

[a] F^A (eq 17) is a lower bound of the free energy, and σ_A (eq 19) is its fluctuation. The results were obtained from samples of $n = 200$ conformations for the smallest bin size, $\delta = \Delta\alpha_k/45$, unit = 5 ps, and all future sample sizes n_f. F^{QH} (see eq 5) and F^{LS} (eq 17 and section II.8) are free energies obtained by the quasi-harmonic approximation and the local states method, respectively, and are based on larger samples (see text). The average potential energy, E_{int}, of the studied samples appears in the bottom row; σ_E is the energy fluctuation (these results are in kcal/mol). All free energies are in kcal/mol and are defined up to the same additive constant for both microstates. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA. The statistical error is defined in footnote *a* of Table 2.

2500, and 5000 conformations. The results for TS^A and $T\Delta S^A$ for the three smallest bin sizes appear in Table 7. For both microstates, the table shows the expected behavior, i.e., that for each bin, S^A decreases as n_f is increased and for a given n_f, S^A decreases as the bin is decreased. The results are not completely converged where the extent of convergence for the free microstate is slightly better than for the bound one. Thus, $T[S^A(\delta,n_f{=}2500) - S^A(\delta,n_f{=}5000)] \sim 0.1$, 0.4, and 0.4 kcal/mol for $\delta = \Delta\alpha_k/15$, $\Delta\alpha_k/30$, and $\Delta\alpha_k/45$, respectively, where the corresponding values for the bound microstate are 0.2, 0.3, and 0.6 kcal/mol. On the other hand, for $n_f = 5000$ $T[S^A(\delta{=}\Delta\alpha_k/30) - S^A(\delta{=}\Delta\alpha_k/45)] \sim 0.2$ for both microstates. As expected, the entropies in solvent are larger than in vacuum, where $TS^A(\Delta\alpha_k/45,n_f{=}5000) = 68.2$ and 68.5 kcal/mol for the free and bound microstates, respectively, in solvent, while the corresponding results in vacuum are $TS^A(\Delta\alpha_k/30,n_f{=}5000) = 65.8$ and 65.5 (the additive constant is assumed to be the same for both environments).

The HSMD results for the entropy are also compared in the table with those obtained using the LS and QH methods, for which larger MD samples (composed of subsamples, see section III.2) of 5000, 8000, and 10 000 conformations were generated (for each microstate) by retaining a conformation every 200 fs (100 MD steps). While both methods are expected to provide overestimations, the QH results for TS are significantly larger than the HSMD values by ~21 and ~24 kcal/mol for the free and bound microstate, respectively, where the LS results (based on $b{=}2$, $l{=}10$) exceed those of QH. These large QH and LS values are also affected by the significantly larger samples used for the QH and LS calculations than for HSMD (see section II.9).

III.7. Entropy Differences in Implicit Water. Table 7 also shows that the results for $T\Delta S^A = T[S^A_{\text{free}} - S^A_{\text{bound}}]$ are converged nicely to -0.4 ± 0.1 kcal/mol for *all* n_f values and bins (even for the not shown $\delta{=}\Delta\alpha_k/10$) (this convergence suggests that decreasing the smallest bin to $\delta{=}\Delta\alpha_k/45$ was not necessary). Thus, in solvent the entropy of the bound microstate is slightly larger than that of the free microstate, unlike in vacuum where this relation is reversed. Again, the QH and LS results for $T\Delta S^A$, −3(1) and −8(1) kcal/mol, respectively differ significantly from the HSMD value.

These perfectly converged results for $T\Delta S^A$ stem from an exact cancellation (see section II.10) of the systematic errors in TS for both microstates, where equal bins $\delta = 100°/l$ are used for the bond angles. This cancellation occurs for a relatively large range of approximations; thus, the (not provided) worst TS^A results for $\delta = \Delta\alpha_k/10$ differ from the best results in Table 7 by $TS^A(\Delta\alpha_k/10,n_f{=}500) - TS^A(\Delta\alpha_k/45,n_f{=}5000) = 3.4$ kcal/mol for both the free and bound microstates; as discussed in section III.4, these differences still constitute lower bounds.

The equal $T\Delta S^A$ results obtained for different n_f values suggest that the level of coverage of both microstates by the future chains during the reconstruction process is comparable and adequate, i.e., the future chains remain within these microstates. To estimate the extent of this coverage, we carried out MD simulations of the *entire* loop generating two samples (for the free and bound microstates) of 1 × 500 (5 ps) conformations and two samples of 10 × 500 [10 × (5 ps)] based on $g = 10$ fs in the same way the future chains are simulated during the reconstruction process (see sections II.6 and III.4 for the loop in vacuum). The $\Delta\alpha_k$ results (eq 13) for the dihedral angles for these samples are presented in Table 6, which shows that in most cases the results for 500 are somewhat smaller and the results for 10 × 500 are larger (but still close) to those of the studied samples; however, several strong deviations from this picture are also observed. Notice that the 500 × 10 results for χ^1(Ser), 188 and 167°, are still significantly smaller than 317 and 321° obtained for the free and bound microstates of the studied samples, respectively. However, the effect of these too small (almost equal) values is expected to get cancelled in entropy differences.

III.8. Free Energy in Implicit Water. Results for the free energy functional, F^A (eq 17), its fluctuation, σ_A (eq 19), and the energies are presented in Table 8. As in vacuum (Table 3), these results are given only for the smallest bin, $\Delta\alpha_k/45$. F^A increases slightly as n_f is increased from 2500 to 5000, while σ_A is unchanged (within the error bars). As expected, the QH and LS results for F underestimate the correct values, and the energy fluctuations are always larger than those for σ_A ($n_f{=}5000$). Finally, the table shows the differences in free energy, ΔF^A, and energy, ΔE, between the free and bound microstates. It is evident that the ΔF^A results are all equal within the statistical errors, and they are also equal to ΔE meaning again that the higher stability (by ~25.5 kcal/mol) of the free microstate over the bound microstate is mostly due to ΔE.

The computer time required in solvent is significantly larger than in vacuum, where reconstructing a structure based on $n_f = 500$ requires ~3.6 h CPU on a 2.1 GHz Athlon processor. This stems from the fact that at each MD step

GB/SA is applied to all of the ~700 atoms, while only the contributions of atoms close to the loop are expected to be affected by the conformational changes of the loop. We have not attempted to reduce the calculation time by eliminating the computation of the (constant) contribution of the atoms remote from the loop.

IV. Summary and Conclusions

As pointed out in section I.7, this study is focused mainly on theoretical and implementation aspects of HSMD as applied (for the first time) to a flexible loop of a protein; for that it has been convenient to treat initially the relatively short loop of pancreatic α-amylase which consists of small residues. The role of this loop in the enzymatic function of pancreatic α-amylase is presently of secondary interest and will be discussed in future studies where the ligand, which interacts with the loop in the bound state (and is missing in the free protein), will be considered, and explicit water—the preferred solvation model—will be introduced.

Still, the relatively large energy (hence free energy) differences between the free and bound microstates (~38 and ~25 kcal/mol in vacuum and solvent, respectively) suggest that the bound microstate would not be visited by the loop in the free protein, i.e., the response of the loop to ligand binding is probably an induced fit rather than a selected fit (see sections I.1, I.7, and II.1). The higher free energy of the bound microstate stems mainly from electrostatic interactions that are contributed by many of the loop atoms (rather a specific one). Thus, while the crystal structures of 1pif and 1pig are similar, they still differ in the structural arrangement of specific side chains, which leads to (relatively) unfavorable electrostatic interactions between the (1pif) template and the bound loop structure (which is superimposed on 1pif). Indeed, the crystal structure of 1pig is significantly better resolved than that of 1pif, where atoms with elevated B-factors (larger than 40) appear in 61 and 153 residues (predominantly charged and polar) of these structures, respectively; also, on average, the B-factors of 1pif are significantly larger than those of 1pig. The unstable MD trajectory obtained initially for the bound microstate is a result of the unfavorable electrostatic interactions, which has led us to generate a bound sample consisting of short trajectories (see section II.1).

In this context we would like to discuss further our result, $S_{\mathrm{free}} \sim S_{\mathrm{bound}}$. Thus, in the presence of the ligand in the active site one would expect $S_{\mathrm{free}} > S_{\mathrm{bound}}$ in accord with the measured B-factors. However, (as discussed above) for both models studied (where the ligand is missing) the energy of the free microstate is significantly lower than that of the bound microstate which suggests that S_{free} should be significantly lower than S_{bound}. The unexpected result, $S_{\mathrm{free}} \sim S_{\mathrm{bound}}$ (also obtained approximately with the quasi-harmonic method) might be attributed to the fact that our bound sample consists of several partial (relatively short) MD samples all starting from the X-ray structure of the bound protein with different sets of velocities, which leads to a relatively concentrated sample of low entropy (see section III.1).

This discussion demonstrates the problems involved in the computational definition of a microstate—a topic that has been ignored to a large extent in the literature but has been given a great deal of thought in this paper. In particular, we have provided strong theoretical arguments that systematic errors in S^{A}(HSMD) for different microstates are comparable and thus get cancelled in differences, ΔS^{A} − our main interest. This means that one can apply highly crude approximations (i.e., small reconstruction samples) decreasing computer time dramatically. Indeed, such cancellation has been observed for peptides[55] and for the loop of α-amylase modeled by the AMBER force field[68] and AMBER with the implicit solvation GB/SA,[69] leading to efficient computations and providing support to our theory. Notice that calculating transition probabilities for different steps k is completely independent, and the reconstruction process is thus completely parallelized. As for peptides, the small statistical errors in $T\Delta S^{\mathrm{A}}$ of 0.1−0.2 kcal/mol is very satisfactory.

An important development has been the realization that the bins can be optimized, leading to improved (i.e., smaller) results for S^{A} hence to reliable results for $T\Delta S^{\mathrm{A}}$ for smaller reconstruction samples. We have not carried out a full bins' optimization but have demonstrated its potential effectiveness by applying to both microstates a uniform set of bins, $\delta =$ constant/l for the bond angles. It is plausible to assume that further bins' optimization would lead to an improved probability density, $\rho^{\mathrm{HS}}(\alpha_K, \cdots, \alpha_1)$ (eq 15), hence to more accurate free energy functionals, F^{D} (eq 22) and F^{B} (eq 21), where the latter exhibits an upper bound behavior. It has also been shown that the contributions of the Jacobian are cancelled out in entropy and free energy differences. The quasi-harmonic approximation and the local states method (as expected) overestimate the entropy but more significantly than for peptides,[53,55] which might reflect strong long-range correlations and anharmonic effects within the loop due to the loop-template interactions.

The theoretical developments introduced in this paper and the conclusions gathered from the application of HSMD to the 7-residue loop of pancreatic α-amylase constitute a mandatory basis for the next step in the development of HSMD—its extension to the same loop modeled by the AMBER force field and explicit water. In this treatment the loop is capped with explicit water,[83] and the entropy is calculated from the sample in a two-stage process, where the loop structure i is reconstructed first leading to S_i (the surrounding waters, which constitute part of the future, are moved as well during the reconstruction of i); next, the water configuration is reconstructed step-by-step in the presence of the frozen loop structure i leading to $S_{\mathrm{w}/i}$. One is interested to estimate $<S_i>$ and $<S_i+S_{\mathrm{w}/i}>$, which constitute measures of flexibility, and their values for the free and bound microstates can be compared (unlike the energy and free energy that due to the ligand depend on different sets of interactions in the free and bound proteins). This study is being carried out presently. After completing these developmental stages HSMD will become a mature tool for studying other flexible loops of interest, such as the 11-residue lid loop of TIM proteins, and problems which require calculating the relative and absolute free energy of binding.

Acknowledgment. We thank Dr. Ron P. White for his contributions in the initial stage of this work. This work was supported by NIH grant 2-R01 GM066090-4 A2.

References

(1) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys*. **1959**, *31*, 459.

(2) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585.

(3) Elber, R.; Karplus, M. *Science* **1987**, *235*, 318.

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

(5) Getzoff, E. D.; Geysen, H. M.; Rodda, S. J.; Alexander, H.; Tainer, J. A.; Lerner, R. A. *Science* **1987**, *235*, 1191.

(6) Rini, J. M.; Schulze-Gahmen, U.; Wilson, I. A. *Science* **1992**, *255*, 959.

(7) Constantine, K. L.; Friedrichs, M. S.; Wittekind, M.; Jamil, H.; Chu, C. H.; Parker, R. A.; Goldfarb, V.; Mueller, L.; Farmer, B. T. *Biochemistry* **1998**, *37*, 7965.

(8) Kessler, H.; Matter, H.; Gemmecker, G.; Kottenhahn, M.; Bates, J. W. *J. Am. Chem. Soc.* **1992**, *114*, 4805.

(9) Baysal, C.; Meirovitch, H. *Biopolymers* **1999**, *50*, 329.

(10) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586.

(11) Eisenmesser, E. Z.; Millet, O.; Labeikovski, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky J. J.; Kay, L. E.; Kern D. *Nature* **2005**, *438*, 117.

(12) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

(13) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.

(14) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395.

(15) Jorgensen, W. L. *Acc Chem. Res*. **1989**, *22*, 184.

(16) Meirovitch, H. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; Vol. 12, p 1.

(17) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047.

(18) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.

(19) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181.

(20) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345.

(21) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249.

(22) Ikeda, K.; Galzitskaya, O. V.; Nakamura, H.; Higo, J. *J. Comput. Chem.* **2003**, *24*, 310.

(23) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795.

(24) Lange, O. F.; Grubmüller, H. *J. Chem. Phys.* **2006**, *124*, 214903.

(25) MacDonald, I. R.; Singer, K. *J. Chem. Phys*. **1967**, *47*, 4766.

(26) Hansen, J.-P.; Verlet, L. *Phys. Rev.* **1969**, *184*, 151.

(27) Hoover, W. G.; Ree, F. H. *J. Chem. Phys*. **1967**, *47*, 4873.

(28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarenden Press: Oxford, 1987.

(29) Kirkwood, J. G. *J. Chem. Phys*. **1935**, *3*, 300.

(30) Zwanzig, R. W. *J. Chem. Phys*. **1954**, *22*, 1420.

(31) Squire, D. R.; Hoover, W. G. *J. Chem. Phys*. **1969**, *50*, 701.

(32) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett*. **1974**, *28*, 578.

(33) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys*. **1977**, *23*, 187.

(34) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(35) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett*. **1989**, *63*, 1195.

(36) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kolmann, P. A. *J. Comput. Chem*. **1995**, *16*, 1339.

(37) Kumar, S.; Payne, P. W.; Vásquez, M. *J. Comput. Chem*. **1996**, *17*, 1269.

(38) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(39) Gō, N.; Scheraga, H. A. *J. Chem. Phys.* **1969**, *51*, 4751.

(40) Gō, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535.

(41) Hagler, A. T.; Stern, P. S.; Sharon, R.; Becker, J. M.; Naider, F. *J. Am. Chem. Soc.* **1979**, *101*, 6842.

(42) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325.

(43) Chang, C. E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017.

(44) Meirovitch, H. *Chem. Phys. Lett.* **1977**, *45*, 389.

(45) Meirovitch, H.; Vásquez, M.; Scheraga, H. A. *Biopolymers* **1987**, *26*, 651.

(46) Meirovitch, H.; Koerber, S. C.; Rivier, J.; Hagler, A. T. *Biopolymers* **1994**, *34*, 815.

(47) Meirovitch, H. *Phys. Rev. A* **1985**, *32*, 3709.

(48) Meirovitch, H.; Scheraga, H. A. *J. Chem. Phys*. **1986**, *84*, 6369.

(49) Meirovitch, H. *J. Chem. Phys.* **2001**, *114*, 3859.

(50) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2004**, *121*, 10889.

(51) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2006**, *124*, 204108.

(52) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2005**, *123*, 214908.

(53) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2005**, *122*, 054903.

(54) Cheluvaraja, S.; Meirovitch, H. *J. Phys. Chem. B* **2005**, *109*, 21963.

(55) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2006**, *125*, 024905.

(56) Qian, M.; Haser, R.; Payan, F. *J. Mol. Biol*. **1993**, *231*, 785.

(57) Qian, M.; Haser, R.; Buisson, G.; Duee, E.; Payan, F. *Biochemistry* **1994**, *33*, 6284.

(58) Qian, M.; Haser, R.; Payan, F. *Protein Sci.* **1995**, *4*, 747.

(59) Machius, M.; Vertesy, L.; Huber, R.; Wiegand, G. *J. Mol. Biol.* **1996**, *260*, 409.

(60) Brayer, G. D.; Sidhu, G.; Maurus, R.; Rydberg, E. H.; Braun, C.; Wang, Y. et al. *Biochemistry* **2000**, *39*, 4778.

(61) Rydberg, E. H.; Li, C.; Maurus, R.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 4492.

(62) Numao, S.; Maurus, R.; Sidhu, G.; Wang, Y.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 215.

(63) Steer, M. L.; Levitzki, A. *FEBS Lett.* **1973**, *31*, 89.

(64) Levitzki, A.; Steer, M. L. *Eur. J. Biochem.* **1974**, *41*, 171.

(65) Aghajari, N.; Feller, G.; Gerday, C.; Haser, R. *Protein Sci.* **2002**, *11*, 1435.

(66) Brayer, G. D.; Luo, Y.; Withers, S. G. *Protein Sci.* **1995**, *4*, 1730.

(67) Ramasubbu, N.; Paloth, V.; Luo, Y.; Brayer, G. D.; Levine, M. J. *Acta Crystallogr., Sect. D: Biol. Crystallogr*. **1996**, *52*, 435.

(68) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

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

(70) Ponder, J. W. *TINKER - software tools for molecular design, version 3.9*; Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 2001.

(71) Meirovitch, H.; Alexandrowicz, Z. *J. Stat. Phys.* **1976**, *15*, 123.

(72) Meirovitch, H. *J. Chem. Phys*. **1999**, *111*, 7215.

(73) Szarecka, A.; White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2003**, *119*, 12084.

(74) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2003**, *119*, 12096.

(75) Meirovitch, H.; Meirovitch, E. *J. Phys. Chem*. **1996**, *100*, 5123.

(76) Meirovitch, H.; Hendrickson, T. F. *Proteins* **1997**, *29*, 127.

(77) Baysal, C.; Meirovitch, H. *Biopolymers* **2000**, *53*, 423.

(78) Meirovitch, H. *J. Chem. Phys*. **1988**, *89*, 2514.

(79) Meirovitch, H.; Vásquez, M.; Scheraga, H. A. *Biopolymers* **1988**, *27*, 1189.

(80) Brady, J.; Karplus, M. *J. Am. Chem. Soc.* **1985**, *107*, 6103.

(81) Gibbs, W. *Elementary Principles in Statistical Mechanics*; Yale University Press: 1902; Chapter XI.

(82) White, R. P.; Meirovitch, H. *J. Chem. Theory Comput.* **2006**, *2*, 1135.

CT700116N

JCTC Journal of Chemical Theory and Computation

Beyond the Wade−Mingos Rules in Bare 10- and 12-Vertex Germanium Clusters: Transition States for Symmetry Breaking Processes

R. B. King,*,[†] I. Silaghi-Dumitrescu,[‡] and M. M. Uţă[‡]

Department of Chemistry, University of Georgia, Athens, Georgia, 30606, and Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, Cluj-Napoca, Roumania

Received August 29, 2007

Abstract: The lowest energy structures of bare Ge_n^z clusters ($n = 10, 12$; $z = -6, 0, +2, +4$) obtained using density functional theory (DFT) at the hybrid B3LYP level often are relatively low-symmetry polyhedra not readily recognizable by the Wade−Mingos rules. However, such optimized structures may arise from higher symmetry transition states through symmetry breaking processes. Thus the lowest energy structures for the Ge_{10}^{6-} and Ge_{12}^{6-} clusters with hyperelectronic *arachno* $2n + 6$ skeletal electron counts are derived from pentagonal and hexagonal prism transition states, respectively, and retain the pentagonal and hexagonal faces of the prisms upon symmetry-breaking optimization. In addition, a variety of capped cube, prism, and antiprism transition states are found for the hypoelectronic Ge_{10}^{4+}, Ge_{12}, and Ge_{12}^{4+} clusters, which go to low-energy low-symmetry optimized structures, typically C_s or C_i, upon following the normal modes of the imaginary vibrational frequencies.

1. Introduction

The Wade−Mingos rules[1−4] historically were derived in order to relate the structures of polyhedral boranes and isoelectronic compounds to the number of skeletal electrons.[5] However, they subsequently have been used to explain the shapes of other cluster structures isoelectronic and isolobal with boranes. According to the Wade−Mingos rules the polyhedra in the so-called *closo* boranes $B_nH_n^{2-}$ and isoelectronic compounds with $2n + 2$ skeletal electrons are the most spherical deltahedra, namely polyhedra in which all faces are triangles and the vertices are as similar as possible. These deltahedral boranes can be considered to be three-dimensional aromatic systems[6,7] with $2n$ of the $2n + 2$ skeletal electrons being used for surface bonding analogous to the σ-bonding in benzene. The remaining two skeletal electrons are used for an n-center two-electron core bond involving overlap of inward pointing radial orbitals from each of the n vertex atoms at the center of the deltahedron. This latter bond in the deltahedral boranes plays an analogous role to the π-bonding in benzene. For the 10- and 12-vertex structures of interest in this paper the most spherical deltahedra found in the *closo* boranes $B_{10}H_{10}^{2-}$ and $B_{12}H_{12}^{2-}$ and related compounds are the D_{4d} bicapped square antiprism and the I_h regular icosahedron, respectively (Figure 1). Note that in counting skeletal electrons in the clusters of interest in this paper either a BH or bare Ge vertex is a donor of two skeletal electrons.

Now consider hyperelectronic (electron-rich) polyhedral boranes having more than $2n + 2$ skeletal electrons. The so-called *nido* boranes with n vertices have $2n + 4$ skeletal electrons and polyhedral structures with one nontriangular face. Frequently such *nido* borane structures can be derived from a *closo* borane structure with $n + 1$ vertices by removing one vertex and its associated edges. Thus the 10-vertex *nido* borane polyhedron, found in the long-known relatively stable[8] $B_{10}H_{14}$, can be obtained by removal of the unique degree 6 vertex from the 11-vertex *closo* deltahedron (Figure 2). Similarly the 12-vertex *nido* borane polyhedron, found in the ligand[9] $C_2B_{10}H_{12}^{2-}$ obtained by reduction of the carborane $C_2B_{10}H_{12}$, can be formally obtained by removal

* Corresponding author e-mail: rbking@chem.uga.edu.

[†] University of Georgia.

[‡] Babeş-Bolyai University.

10.1021/ct7002226 CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/15/2007

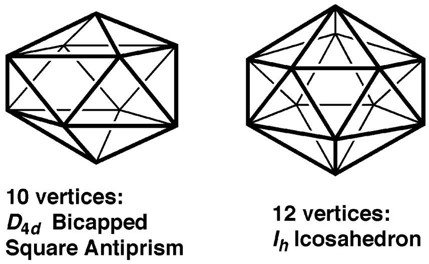


Figure 1. The most spherical (*closo*) deltahedra with 10 and 12 vertices.

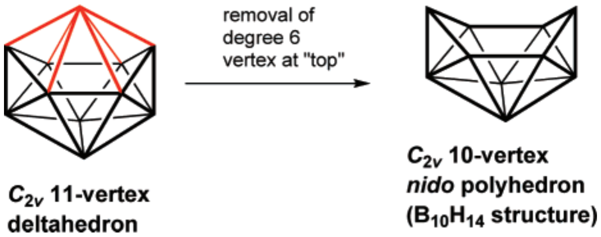


Figure 2. Conversion of the 11-vertex most spherical deltahedron to the 10-vertex *nido* polyhedron (found in $B_{10}H_{14}$) by removal of the unique degree 6 vertex (the "top" vertex) and the associated edges (colored red).

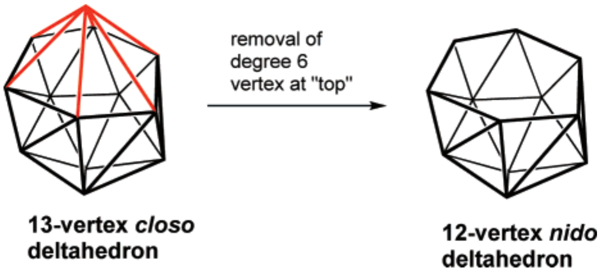


Figure 3. Conversion of the 13-vertex most spherical deltahedron to the 12-vertex *nido* polyhedron (found in the $C_2B_{10}H_{12}^{2-}$ ligand in (η^5-C_5H_5)$CoC_2B_{10}H_{12}$) by removal of the unique degree 6 vertex (the "top" vertex) and the associated edges (colored red).

of a degree 6 vertex from the 13-vertex *closo* polyhedron found in metallaboranes such as (η^5-C_5H_5)$CoC_2B_{10}H_{12}$ (Figure 3).

The Wade−Mingos rules in borane chemistry have been extended to systems even more hyperelectronic than the *nido* compounds such as the *arachno* compounds with $2n + 6$ skeletal electrons and two nontriangular faces or one large nontriangular face and the *hypho* compounds with $2n + 8$ skeletal electrons and an even more open structure. In principle, the *arachno* and *hypho* structures with n vertices can be derived from *closo* structures with $n + 2$ or $n + 3$ vertices, respectively, by removal of two or three vertices, respectively. However, as the structures become electron-richer the increasingly open polyhedra become increasingly less recognizable.

The Wade−Mingos rules[1−4] are more difficult to apply to hypoelectronic (electron-poor) clusters containing fewer than the $2n + 2$ skeletal electrons of *closo* structures. Such systems are not found in borane and carborane derivatives containing exclusively boron and carbon vertices so that they were not considered in Wade's original work.[1,2] However, hypoelectronic structures are found in isoelectronic metal carbonyl clusters and bare post-transition element clusters. Hypoelectronic structure types found in systems with n vertices and less than $2n + 2$ skeletal electrons include the following: (1) a capped deltahedron with $m < n$ vertices, typically for a system with $m + 2$ skeletal electrons such as the capped octahedral osmium carbonyl cluster[10] $Os_7(CO)_{21}$ ($v = 7$ but $m = 6 \Longrightarrow 14$ skeletal electrons) and (2) "flattened" deltahedra with f "flattened" vertices pushed toward the center of the deltahedron,[11] typically for a system with $v - f + 2$ skeletal electrons such as the In_{11}^{7-} cluster found in K_8In_{11} ($v = 11$, $f = 3 \Longrightarrow 18$ skeletal electrons).[12]

In recent years the chemistry of bare post-transition-metal clusters has expanded greatly from the original work of Zintl and co-workers.[13−16] Such clusters can be considered to be formally isoelectronic with boranes and carboranes.[17] Thus a bare group 14 element vertex (Si, Ge, Sn, Pb) is a donor of two skeletal electrons like a B−H vertex in boranes. Similarly a bare group 15 element vertex (P, As, Sb, Bi) is a donor of three skeletal electrons like a C−H vertex in carboranes. However, in many cases the polyhedra found in bare post-transition-metal clusters are different from those found in boranes and related compounds. Furthermore, they do not relate obviously to polyhedra suggested by the Wade−Mingos rules,[1−4] particularly in the cases of electron-poor clusters containing bare group 13 elements (Al, Ga, In, Tl), which are donors of only one skeletal electron. In order to understand such unusual polyhedra and the chemical bonding in such structures we have performed density functional theory (DFT) studies of germanium clusters containing from 5 to 12 germanium atoms.[18−23] Germanium was chosen as a model vertex atom to minimize the charges on clusters isoelectronic with the known molecules of interest.

Our studies as well as the work of others have indicated major differences between isoelectronic boranes and carboranes, on the one hand, and bare germanium and other post-transition-metal clusters, on the other hand. Examples are the following: (1) The antiaromaticity of the icosahedral E_{12}^{2-} (E = Si, Ge) as compared with the strong aromaticity of the isoelectronic $B_{12}H_{12}^{2-}$ as noted above.[24,25] (2) The lowest energy structure for Ge_{11}^{2-} is not the most spherical 11-vertex deltahedron[22] found in the stable borane $B_{11}H_{11}^{2-}$. (3) The lowest energy structure for Ge_8^{2-} is the spherically aromatic T_d tetracapped tetrahedron rather than the bisdisphenoid found in $B_8H_8^{2-}$ and related compounds.[19] These differences between isoelectronic bare germanium and borane clusters appear to be a consequence of the fact that the external germanium lone pair electrons can participate in the skeletal bonding, whereas no comparable electrons are available from the B−H vertices of boranes.

We have also observed many low-energy low-symmetry (not readily recognizable) germanium clusters Ge_n^z, particularly those with fewer than $2n + 2$ skeletal electrons. Such polyhedra often arise during the optimization of more obvious symmetrical polyhedra by following imaginary vibrational frequencies. The more symmetrical and thus more recognizable polyhedra can then be considered as transition states linking isomeric polyhedra of low symmetry. Thus a useful way of characterizing unsymmetrical low-energy

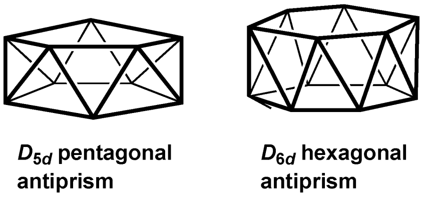


Figure 4. The pentagonal and hexagonal antiprisms as possible *arachno* polyhedra with two nontriangular faces (namely pentagons and hexagons, respectively).

polyhedra is by the more symmetrical transition states from which they arise.

A previous paper from our group[26] considers low-energy low-symmetry structures of hypoelectronic 11-vertex bare germanium clusters Ge$_{11}^{z}$. In that case, consideration of transition states leading to such clusters is less useful since relatively few chemically relevant 11-vertex polyhedra are readily recognizable because of their generally low symmetries. The present paper discusses the transition states leading to the low-energy low-symmetry structures of 10- and 12-vertex bare germanium clusters Ge$_{10}^{z}$ $z = +4$ and Ge$_{12}^{z}$ $z = 0, +4$ found in our previous work.[21,23] In this case, the low-symmetry structures often arise from readily recognizable symmetrical transition states. This paper characterizes such transition states, both for hypoelectronic Ge$_{10}^{z}$ and Ge$_{12}^{z}$ clusters as well as for hyperelectronic Ge$_{10}^{6-}$ and Ge$_{12}^{6-}$ clusters. The Ge$_{10}^{6-}$ and Ge$_{12}^{6-}$ clusters with the $2n + 6$ *arachno* skeletal electron counts are of interest since the readily recognizable pentagonal and hexagonal antiprism structures satisfying the *arachno* requirement of $2n + 6$ skeletal electrons and two obvious nontriangular faces (Figure 4) are not the lowest energy structures.

2. Theoretical Methods

The density functional theory (DFT) methods used in this paper are described in our previous papers on Ge$_{10}^{z}$ and Ge$_{12}^{z}$ clusters.[21,23] Thus the geometry optimizations were carried out at the hybrid DFT B3LYP level[27−30] with the 6-31G(d) (valence) double-ζ quality basis functions extended by adding one set of polarization (d) functions. The Gaussian 98 package of programs[31] was used in which the fine grid (75, 302) is the default for numerically evaluating the integrals and the tight (10^{-8}) hartree stands as default for the self-consistent field convergence. The symmetries were maintained during the initial geometry optimization processes. In the systems of interest in this paper the transition states of interest were optimized structures with significant imaginary vibrational frequencies, typically above 100*i* cm^{-1}. Symmetry breaking using the normal modes of these transition states defined by these imaginary vibrational frequencies was then used to determine optimized structures with minimum energies. Both the transition states and the final optimized structures are discussed in this paper with their relationships being depicted in Figures 5−11 (and Figures 1S−6S in the Supporting Information) with the energy differences between the transition state and the final structure approximately according to scale.

One might raise the legitimate question if this method is suitable for describing subtle electronic effects governing the structure of germanium clusters. In this connection Archibong and St. Amant[32] have shown that CCSD(T) calculations on Ge$_{6}^{z}$ ($z = 0, -1$) give similar results to those obtained at the B3LYP DFT level of theory.

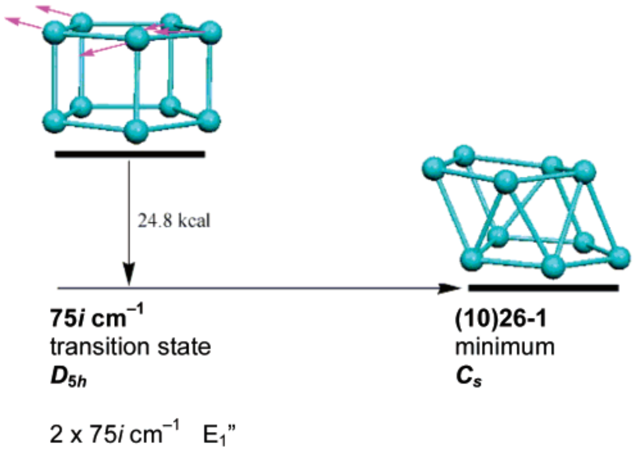


Figure 5. The distortion of the pentagonal prism transition state in Ge$_{10}^{6-}$ along the E$_1$ normal mode of the 75*i* cm^{-1} vibrational frequency to give the global minimum **(10)26-1**.

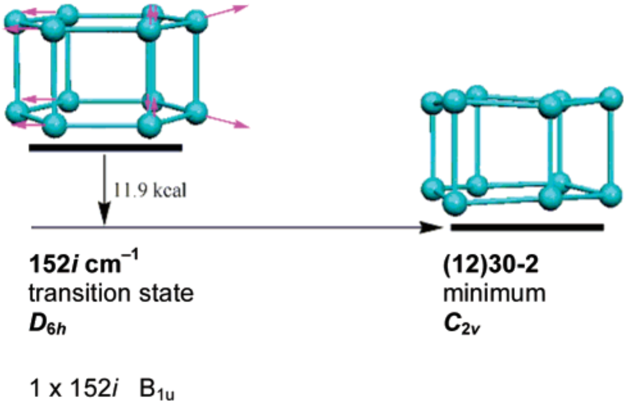


Figure 6. The distortion of the hexagonal prism transition state in Ge$_{12}^{6-}$ along the B$_{1u}$ normal mode of the 152*i* cm^{-1} vibrational frequency to give the lowest energy polyhedral structure **(12)30-2**.

The individual structures are labeled according to the number of skeletal electrons and relative energies with the specific relative energy designations matching those in the previous papers.[21,23] In addition, structures with 10 and 12 vertices are distinguished by the designations (10) and (12), respectively, in front of their structure labels. Thus the lowest energy structure of the 10-vertex 26 skeletal electron system Ge$_{10}^{6-}$, designated as **26-1** in the previous paper,[21] is now designated as **(10)26-1** (see Figure 5) in order to differentiate it from the lowest energy 12-vertex structure with 26 skeletal electrons, which would now be designated as **(12)26-1**.

3. Results

3.1. The Systems Ge$_{10}^{6-}$ and Ge$_{12}^{6-}$ with the *arachno* 2*n* + 6 Skeletal Electron Counts. The Wade−Mingos rules[1−4] suggest that the D_{5d} pentagonal antiprism (Figure 4) should be the global minimum for the 26 ($=2n + 6$ for $n = 10$) skeletal electron 10-vertex system Ge$_{10}^{6-}$ with an *arachno* skeletal electron count. Indeed a pentagonal antiprismatic

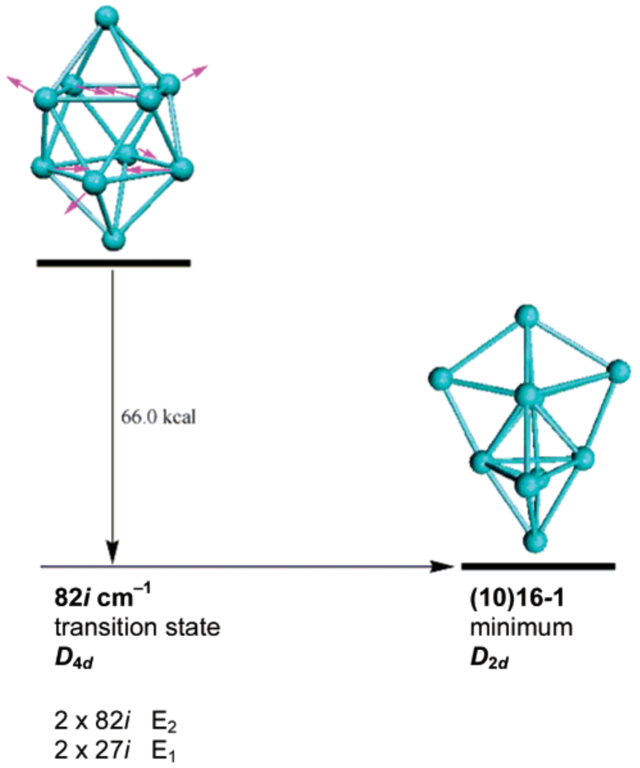


Figure 7. The distortion of the D_{4d} bicapped square antiprism transition state in Ge_{10}^{4+} along the E_2 normal mode of the 82*i* cm^{-1} vibrational frequency to give the lowest energy polyhedral structure **(10)16-1**.

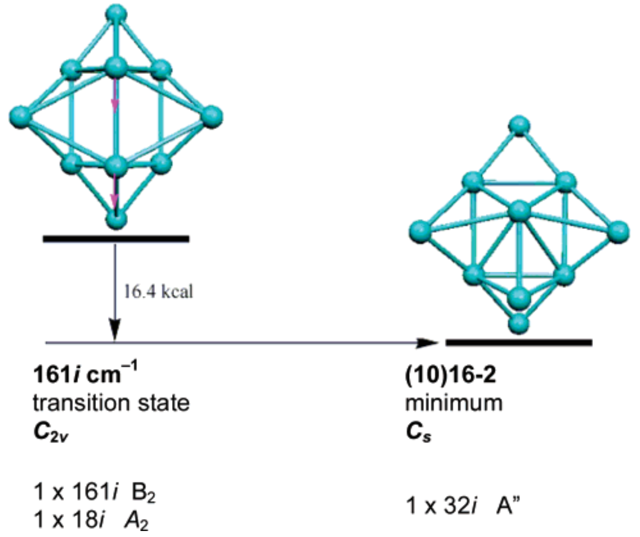


Figure 8. The distortion of the C_{2v} tetracapped trigonal prism transition state in Ge_{10}^{4+} along the B_2 normal mode of the 161*i* vibrational frequency to give the C_s polyhedral structure **(10)16-2**.

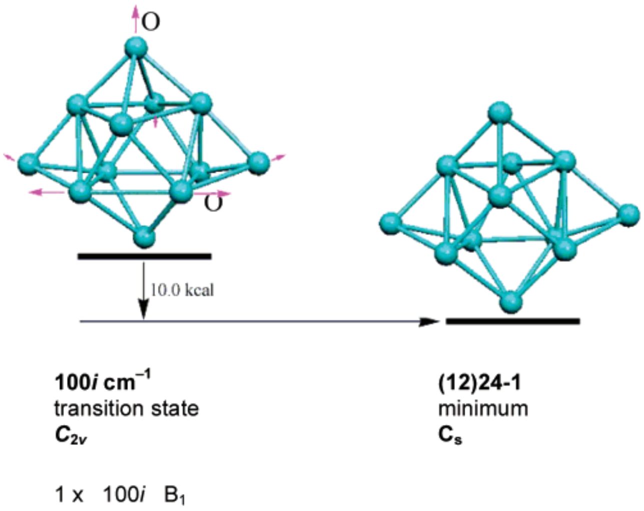


Figure 9. The distortion of the C_{2v} tetracapped square antiprism transition state in Ge_{12} along the normal mode of the 100*i* B_1 vibrational frequency to give the C_s polyhedral structure **(12)24-1**.

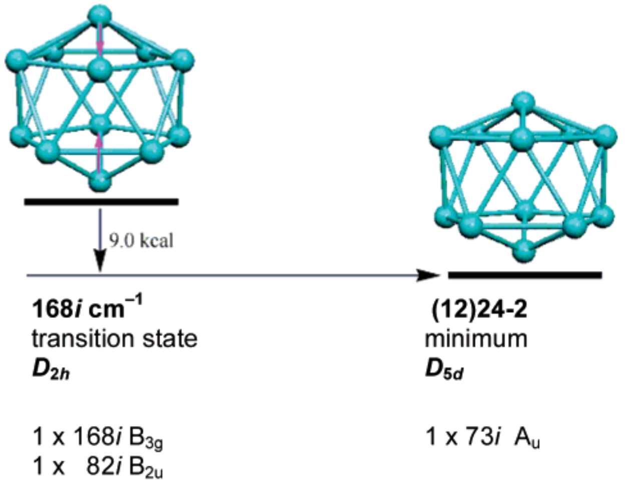


Figure 10. The distortion of the D_{2h} irregular icosahedron transition state in Ge_{12} along the B_{2g} normal mode of the 168*i* vibrational frequency to give the D_{5d} bicapped pentagonal antiprismatic structure **(12)24-2**.

structure **(10)26-3** is found for Ge_{10}^{6-} but at 17.1 kcal/mol above the global minimum[21] **(10)26-3**. Optimization of a D_{5h} pentagonal prism starting structure leads to a pentagonal prismatic transition state with a 75*i* cm^{-1} imaginary vibrational mode. Following the corresponding E_1'' normal mode lowers the energy of the structure by 24.8 kcal/mol leading to the global minimum **(10)26-1** observed[21] for Ge_{10}^{6-} (Figure 5). The polyhedron in **(10)26-1** is derived from a D_{5h} pentagonal prism by distortion of the top pentagonal face relative to the bottom face so that the five rectangular faces linking the two pentagonal faces in the original prism become two rectangular and six triangular faces by the addition of diagonals across three of the original rectangular faces. This distortion necessarily destroys the original C_5 axis in the pentagonal prism. Also note that the **(10)26-1** polyhedron is intermediate between the pentagonal prism with five rectangular faces between the two pentagonal faces and the pentagonal antiprism with ten triangular faces between the two pentagonal faces.

The Wade−Mingos rules[1−4] also suggest that the D_{6d} hexagonal antiprism (Figure 4) should be the global minimum for the 30 ($=2n + 6$ for $n = 12$) skeletal electron 12-vertex system Ge_{12}^{6-} with the *arachno* skeletal electron count. Indeed a hexagonal antiprismatic structure **(12)30-5** is found for Ge_{12}^{6-} but at 23.1 kcal/mol above the lowest energy polyhedral structure **(12)30-2.**[23] Optimization of the D_{6h} hexagonal prism leads to a hexagonal prismatic transition

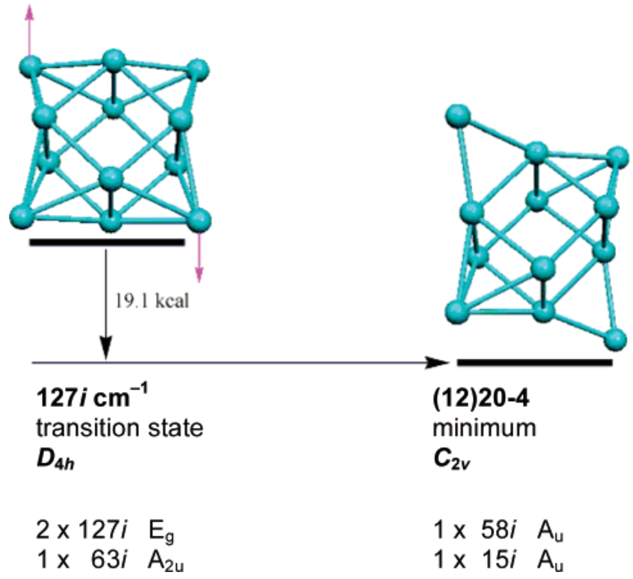


Figure 11. The distortion of the D_{4h} tetracapped cube transition state in Ge_{12}^{4+} along the E_g normal mode of the $127i$ vibrational frequency to give the C_{2v} structure **(12)20-4**.

state with a $152i$ cm^{-1} imaginary vibrational mode. Following the corresponding B_1 normal mode lowers the energy of the structure by 11.9 kcal/mol leading to the lowest energy polyhedral structure **(12)30-2** calculated for Ge_{10}^{6-} (Figure 6). This structure retains the topology of the hexagonal prism (i.e., the two hexagonal faces and the six quadrilateral faces between the two hexagonal faces) but destroys the C_6 axis.

3.2. The Hypoelectronic Cluster Ge_{10}^{4+}. The global minimum **(10)16-1** for Ge_{10}^{4+} with 16 skeletal electrons is derived from a relatively high energy D_{4d} bicapped square antiprism transition state (Figure 7). Following the E_2 normal mode corresponding to the largest imaginary vibrational frequency ($82i$) reduces the energy of the system by a gigantic 66.0 kcal/mol. The symmetry is concurrently reduced from D_{4d} to D_{2d} leading to a polyhedron derived from two interlocking planar pentagons at right angles to each other, which resembles a “Siamese twin” of two pentagonal bipyramids (Figure 7). Note that the pentagonal bipyramid building block of structure **(10)16-1** is the most spherical deltahedron with seven vertices and thus requires $2n + 2 = 16$ skeletal electrons for $n = 7$ by the Wade−Mingos rules.[1−4]

The next higher energy structure **(10)16-2** for Ge_{10}^{4+}, at 14.5 kcal/mol above **(10)16-1** discussed above, is derived from a C_{2v} tetracapped trigonal prism transition state (Figure 8). Following the B_2 normal mode of the largest imaginary frequency ($161i$) adds diagonals across two rectangular faces of the underlying trigonal prism with concurrent conversion of the corresponding two rectangular face caps to triangular face caps. The symmetry is thus reduced from C_{2v} to C_s and the energy by 16.4 kcal/mol (Figure 8). A relatively small imaginary vibrational frequency of $32i$ remains in **(10)16-2**, which might arise from a numerical integration error.[33,34]

A C_{2v} bicapped cube is the transition state to a C_s higher energy structure **(10)16-4** for Ge_{10}^{4+} (Figure 1S), which lies 20.6 kcal/mol above the global minimum **(10)16-1** (Figure 7). Following the B_1 normal mode corresponding to the relatively large $342i$ imaginary vibrational frequency in the bicapped cube reduces the symmetry from C_{2v} to C_s. However, the energy is reduced by only 7.9 kcal/mol with no significant change in the polyhedron topology.

3.3. The Neutral 12-Vertex Cluster Ge_{12}. The lowest energy structures for the neutral Ge_{12} cluster with 24 skeletal electrons are relatively low-symmetry structures derived by distortion of higher energy transition states. The global minimum for Ge_{12}, namely **(12)24-1** (Figure 9), is derived from a C_{2v} tetracapped square antiprism transition state by distortion along the B_1 normal mode corresponding to the $100i$ cm^{-1} vibrational frequency. The symmetry is reduced from C_{2v} to C_s, and two pairs of triangular faces in the original tetracapped square antiprism are opened (following the arrows marked by “O” in Figure 9) up into two quadrilateral faces. Removal of the two vertices capping triangular faces from the **(12)24-1** polyhedron leaves a 10-vertex polyhedron with 14 triangular faces and one quadrilateral face. This is a 10-vertex *nido* polyhedron and thus expected by the Wade−Mingos rules[1−4] to have the $2n + 4 = 24$ for $n = 10$ skeletal electrons found in Ge_{12}.

The next higher energy structure **(12)24-2** for Ge_{12}, at 21.9 kcal/mol above **(12)24-1** discussed above, is derived from an irregular D_{2h} icosahedron transition state by distortion along the B_{3g} normal mode corresponding to the largest imaginary frequency at $168i$ cm^{-1} (Figure 10). This normal mode involves compression of a pair of antipodal vertices and leads to a compressed (oblate) bicapped pentagonal antiprism still retaining the topology of the original icosahedron (i.e., no edges are broken). This process lowers the energy by 9.0 kcal/mol (Figure 10).

The next higher energy structure **(12)24-3** for Ge_{12}, at 23.5 kcal/mol above **(12)24-1** discussed above, is derived from a D_{6d} hexagonal antiprism transition state by distortion along the E_4 normal mode corresponding to the largest imaginary frequency at $143i$ cm^{-1} (Figure 2S). This process lowers the energy by a very large 58.9 kcal/mol and the symmetry from D_{6d} to D_{2d}. Each of the hexagonal faces of the original hexagonal antiprism becomes a pair of trapezoidal faces by formation of new transannular edges.

A higher energy structure for Ge_{12}, namely **(12)24-6** at 28.2 kcal/mol above the global minimum **(12)24-1**, is obtained from a D_{4h} tetracapped cube by distortion along the E_g normal mode corresponding to the $114i$ cm^{-1} vibrational frequency (Figure 3S). The process lowers the energy by 12.8 kcal/mol and the symmetry from D_{4h} to C_{2v}. The polyhedron in the resulting structure **(12)24-6** can be described as a pair of C_{2v} bicapped trigonal prisms sharing their uncapped rectangular faces.

3.4. The Lowest Energy Structure of Ge_{12}^{2+}. The lowest energy structure for Ge_{12}^{2+}, namely the D_{3d} hexacapped octahedron **(12)22-1**, is derived from a D_{4h} tetracapped cube transition state (Figure 4S) by a distortion along the E_g normal mode corresponding to the $95i$ cm^{-1} vibrational frequency. This process replaces the original 4-fold axis with a 3-fold axis and lowers the energy by 42.8 kcal/mol. Since the D_{3d} point group of the hexacapped octahedron is obviously not a subgroup of the D_{4h} point group of the

Table 1. Seven Transition States within 25 kcal/mol of the Final Optimized Structures

figure	species	transition state	imaginary frequencies[a]	symmetry breaking[c]	energy gain[b]/ binding energy/atom[d] (transition state/minimum)
			Hyperelectronic (*arachno*) Species		
5	Ge_{10}^{6-}	pentagonal prism	$75i(E_1'')$	$D_{5h}\rightarrow C_s$	24.8/+8.1/5.6
6	Ge_{12}^{6-}	hexagonal prism	$152i(B_{1u})$	$D_{6h}\rightarrow C_{2v}$	11.9/−19.9/−20.9
			Hypoelectronic Species		
8	Ge_{10}^{4+}	tetracapped trigonal prism	$161i(B_2),18i(A_2)$	$C_{2v}\rightarrow C_s$	16.4/−55.9/−57.5
S1	Ge_{10}^{4+}	bicapped cube	$342i(B_1),48i(B_2)$	$C_{2v}\rightarrow C_s$	7.9/−56.1/−56.9
9	Ge_{12}	tetracapped square antiprism	$100i(B_1)$	$C_{2v}\rightarrow C_s$	10.0/−112.2/−113.0
10	Ge_{12}	irregular icosahedron (D_{2h})	$168i(B_{3g}),82i$	$D_{2h}\rightarrow D_{5d}(73i)^c$	9.0/−110.4/−111.2
11	Ge_{12}^{4+}	tetracapped cube	$127i(E_g),62i(A_{2u})$	$D_{4h}\rightarrow C_{2v}(58i)$	19.1/−69.8/−71.4

[a] Residual imaginary vibrational frequencies are given in parentheses. [b] Difference between the energy of the transition state and that of the optimized structure (kcal/mol). [c] This represents a symmetry change rather than symmetry breaking. [d] Difference between the energy of the cluster and the sum of the energies of the components divided by the number of atoms (kcal/mol).

tetracapped cube transition state, this is not a simple symmetry breaking process.

3.5. Several Structures for Ge_{12}^{4+}. The global minimum for Ge_{12}^{4+}, namely **(12)20-1**, is a C_{2v} structure that does not resemble anything actually found in a chemical system.[23] The next highest energy structure for Ge_{12}^{4+}, namely **(12)-20-2** at 8.0 kcal/mol above the global minimum **(12)20-1**, is a D_{4h} double cube, which can be derived from an O_h cuboctahedron transition state by distortion along the E_u normal mode of the highest imaginary vibration, namely the 417*i* cm^{-1} frequency (Figure 5S). This process is a drastic rearrangement that lowers the energy of the system by a gigantic 164.4 kcal/mol.

The next higher energy structure for Ge_{12}^{4+}, namely **(12)-20-3** at 8.9 kcal/mol above **(12)20-1**, is derived from a D_{4h} double square antiprism transition state. Following the E_g normal mode corresponding to the highest imaginary vibrational frequency, namely the 123*i* cm^{-1} frequency, lowers the energy by a large 62.0 kcal/mol and the symmetry from D_{4h} to C_i to give a much more open structure than the original double square antiprism (Figure 5S).

The next structure for Ge_{12}^{4+}, namely the C_{2v} structure (**12**)**20-4** at 10.5 kcal/mol above the global minimum **(12)-20-1**, is derived from a D_{4h} tetracapped cube transition state (Figure 11). Distortion along the E_g normal mode of the largest imaginary frequency at 127*i* reduces the symmetry from D_{4h} to C_{2v} with two opposite face caps becoming edge caps. This process results in an energy gain of 19.1 kcal/mol.

4. Discussion

The figures in the text and the Supporting Information depict 12 transition states leading to previously obtained[21,23] optimized structures for Ge_n^{6-} ($n = 10, 12$), Ge_{10}^{4+}, Ge_{12}, Ge_{12}^{2+}, and Ge_{12}^{4+} including global minima. The polyhedra in most of these optimized low-energy structures are not readily recognizable by the Wade−Mingos rules.[1−4] Of these 12 transition states, seven of them (Table 1) are within 25 kcal/mol of the optimized structure and thus are potentially chemically significant. The remaining five higher energy transition states, at energies ranging from 42.8 to 164.4 kcal/mol above the corresponding optimized structure, are significant mainly in indicating the route used for the DFT optimizations. These are the closest ones to the final optimized structures of any of the initial structures investigated. This is particularly true for the D_{4h} tetracapped cube transition state leading to the D_{3d} hexacapped octahedron global minimum **(12)22-1** of Ge_{12}^{4+} (Figure S3), since D_{3d} is not a subgroup of D_{4h} so that the conversion of the tetracapped cube to a hexacapped octahedron is not a simple symmetry breaking process.

A true transition state exhibits exactly one significant imaginary vibrational frequency, the normal mode of which indicates the pathway to the corresponding minimum point. Additional small imaginary frequencies significantly below 100*i* cm^{-1} can be attributed to errors arising from the numerical integration procedure.[33,34] The seven transition states listed in Table 1 all have exactly one imaginary vibrational frequency at or above 100*i* cm^{-1} except for the pentagonal prism transition state for Ge_{10}^{6-} (Figure 5), which has a single imaginary vibrational frequency at 75*i* cm^{-1}, which was followed in the optimization to the global minimum **(10)26-1** for Ge_{10}^{6-}. Two of the optimized structures in Table 1, namely the D_{5d} bicapped pentagonal antiprism structure **(12)24-2** for Ge_{12}, derived from a D_{2h} irregular icosahedron transition state (Figure 11), and the C_{2v} structure **(12)20-4** for Ge_{12}^{4+}, derived from a D_{4h} tetracapped cube transition state (Figure 11), have residual imaginary vibrational frequencies below 100*i* cm^{-1} after optimization. These structures are low-energy structures for Ge_{12} and Ge_{12}^{4+}, respectively, but are not their global minima.

Most of the high-energy transition states correspond to structures that are relatively far removed from the final structure and have more than one imaginary vibrational frequency large enough to be significant, thereby corresponding more accurately to higher order saddle points. The most extreme case of this type found in this work is the cuboctahedron for Ge_{12}^{4+}, which has three imaginary vibrational frequencies above 100 cm^{-1}, namely 417*i* cm^{-1} (E_u), 290*i* cm^{-1} (T_{2u}), and 115*i* (T_{1g}) as well as smaller imaginary vibrational frequencies at 36*i* (A_{2g}) and 29*i* (E_g). Following the normal mode corresponding to the 417*i* cm^{-1} vibrational frequency results in the gigantic energy lowering of 164.4 kcal/mol and a rather drastic rearrangement leading to a D_{4h} double cube stationary point **(12)20-2**. This corresponds to the lowest energy chemically realistic structure[23] found for Ge_{12}^{4+}.

In general the binding energies per atom are higher for the neutral clusters and decrease both in the negatively and positively charged species as expected. Notable is also the finding that the binding energies for the minima have only slightly lower values than those of the corresponding transition states.

Acknowledgment. We are indebted to the National Science Foundation for partial support of this work under Grants CHE-0209857 and CHE-0716718. Part of this work was undertaken with the financial support from CNCSIS-Roumania.

Supporting Information Available: Figures 1S−6S as cited in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wade, K. *Chem. Commun. (Cambridge)* **1971**, 792.

(2) Wade, K. *Adv. Inorg. Chem. Radiochem*. **1976**, *18*, 1.

(3) Mingos, D. M. P. *Nat. Phys. Sci*. **1972**, *99*, 236.

(4) Mingos, D. M. P. *Acc. Chem. Res*. **1984**, *17*, 311.

(5) Williams, R. E. *Chem. Rev*. **1992**, *92*, 177.

(6) King, R. B.; Rouvray, D. H. *J. Am. Chem. Soc*. **1977**, *99*, 7834.

(7) King, R. B. *Chem. Rev*. **1991**, *101*, 1119.

(8) Kasper, J. S.; Lucht, C. M.; Harker, D. *Acta Crystallogr*. **1950**, *3*, 436.

(9) Dustin, D. F.; Dunks, G. B.; Hawthorne, M. F. *J. Am. Chem. Soc*. **1973**, *95*, 1109.

(10) Eady, C. R.; Johnson, B. F. G.; Lewis, J.; Mason, R.; Hitchcock, P. B.; Thomas, K. M. *Chem. Commun. (Cambridge)* **1977**, 385.

(11) King, R. B. *Rev. Roum. Chim.* **2002**, *47*, 1005.

(12) Sevov, S. C.; Corbett, J. D. *Inorg. Chem.* **1991**, *30*, 4875.

(13) Zintl, E.; Goubeau, J.; Dullenkopf, W. *Z. Phys. Chem., Abt. A* **1931**, *154*, 1.

(14) Zintl, E.; Harder, A. *Z. Phys. Chem., Abt. A* **1931**, *154*, 47.

(15) Zintl, E.; Dullenkopf, W. *Z. Phys. Chem., Abt. B* **1932**, *16*, 183.

(16) Zintl, E.; Kaiser, H. *Z. Anorg. Allgem. Chem.* **1933**, *211*, 113.

(17) King, R. B. *Inorg. Chim. Acta* **1982**, *57*, 79.

(18) Ge_5, Ge_6, and Ge_7: King, R. B.; Silaghi-Dumitrescu, I.; Kun, A. *J. Chem. Soc., Dalton Trans.* **2002**, 3999.

(19) Ge_8: King, R. B.; Silaghi-Dumitrescu, I.; Lupan, A. *Dalton Trans.* **2005**, 1858.

(20) Ge_9: King, R. B.; Silaghi-Dumitrescu, I. *Inorg. Chem.* **2003**, *42*, 6701.

(21) Ge_{10}: King, R. B.; Silaghi-Dumitrescu, I.; Uţă, M. M. *Inorg. Chem.* **2006**, *45*, 4974.

(22) Ge_{11}: King, R. B.; Silaghi-Dumitrescu, I.; Lupan, A. *Inorg. Chem.* **2005**, *44*, 3579.

(23) Ge_{12}: King, R. B.; Silaghi-Dumitrescu, I.; Uţă, M. M. *Dalton Trans.* **2007**, 364.

(24) King, R. B.; Heine, T.; Corminboeuf, C.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2004**, *126*, 430.

(25) Chen, Z.; Neukemans, S.; Wang, X.; Janssens, E.; Zhou, Z.; Silverans, R. E.; King, R. B.; Schleyer, P. v. R.; Lievens, P. *J. Am. Chem. Soc.* **2006**, *128*, 12829.

(26) King, R. B.; Silaghi-Dumitrescu, I.; Lupan, A. *Chem. Phys.* **2006**, *327*, 344.

(27) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys*. **1980**, *58*, 1200.

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

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

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

(31) 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.; Rega, N.; 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 98, Revision A.11.3*; Gaussian, Inc.: Pittsburgh, PA, 2002.

(32) Archibong, E. F.; St-Amant, A. *J. Chem. Phys.* **1998**, *109*, 962.

(33) Xie, Y.; Schaefer, H. F.; King, R. B. *J. Am. Chem. Soc.* **2000**, *122*, 8746.

(34) Papas, B. N.; Schaefer, H. F. *J. Mol. Struct. THEOCHEM* **2006**, *768*, 175.

CT7002226

JCTC Journal of Chemical Theory and Computation

ERRATUM

Minimalist Explicit Solvation Models for Surface Loops in Proteins. [J. Chem. Theory Comput. 2008, 4, 1] [*J. Chem. Theory Comput. 2,* 1135−1151 (2006)]. By Ronald P. White and Hagai Meirovitch*. Department of Computational Biology, University of Pittsburgh School of Medicine, 3064 Biomedical Tower 3, Pittsburgh, Pennsylvania 15260

Page 1150. Before the last sentence of the paper (starting "One such method...") should appear the following two sentences that have been omitted by mistake:

The calculation of such extensive variables (e.g., *E*, *F*, and *S*) with enough accuracy is possible because of the relatively small size of the system. For example, the average energy (over the five trajectories) with $N_W = 120$ ($R_{cap} = 18$ Å) is -1587.4 ± 0.6, -1505.2 ± 0.7, -1894.9 ± 8.3, and -1727.9 ± 0.6 kcal/mol for the loop of RNase, ser-proteinase, and loops 1 and 2 of proteinase, where $R_{temp} = 15$, 13, 13, and 13 Å, respectively.

CT600317D

10.1021/ct600317d
Published on Web 11/18/2006