JCTC Journal of Chemical Theory and Computation

# Active Participation of the $Mg^{2+}$ Ion in the Reaction Coordinate of RNA Self-Cleavage Catalyzed by the Hammerhead Ribozyme

Kin-Yiu Wong,[†] Tai-Sung Lee,[‡] and Darrin M. York*[,†,‡]

*Department of Chemistry, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota 55455, United States, and BioMaPS Institute and Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Rd., Piscataway, New Jersey 08854-8087, United States*

Received August 24, 2010

**Abstract:** We report results from combined quantum mechanical/molecular mechanical (QM/MM) free energy simulations to explore metal-assisted phosphoryl transfer and general acid catalysis in the extended hammerhead ribozyme. The mechanisms considered here assume that the 2′OH group of C17 has already been activated (i.e., is deprotonated) and acts as a nucleophile to go on an in-line attack to the adjacent scissile phosphate, passing through a pentavalent phosphorane intermediate/transition state, followed by acid-catalyzed departure of the O5′ leaving group of C1.1. A series of six two-dimensional potential of mean force profiles are reported in this study, requiring an aggregate of over 100 ns of QM/MM simulation. The simulations employ the AM1/d-PhoT semiempirical quantum model and linear-scaling QM/MM-Ewald method and explore mechanistic pathways for the self-cleavage. Results support the plausibility of a cleavage mechanism where phosphoryl transfer and general acid catalysis are stepwise, and where the catalytic divalent metal ion plays an active role in the chemical steps of catalysis.

Small self-cleaving ribozymes such as the hammerhead ribozyme (HHR) have been instrumental as model systems for RNA catalysis.[1−3] Recently, an extended HHR structure was determined by X-ray crystallography at 2.0 Å resolution,[4] in which a divalent metal ion was observed near the active site. Subsequent computer simulations lend credence to the possibility that this metal ion may play an active role in catalysis,[4,5] although free energy profiles to elucidate specific pathways have not yet been reported.

We report results from combined quantum mechanical/molecular mechanical (QM/MM) free energy simulations to explore metal-assisted phosphoryl transfer and general acid catalysis in the extended HHR. The mechanisms considered here assume that the 2′OH group of C17 has already been activated (i.e., is deprotonated) and acts as a nucleophile to go on an in-line attack to the adjacent scissile phosphate, passing through a pentavalent phosphorane intermediate/transition state, followed by acid-catalyzed departure of the O5′ leaving group of C1.1. The general acid is assumed to be the 2′OH group of G8.[6−8]

A series of six 2D potential of mean force (PMF) profiles is herein reported, requiring an aggregate of over 100 ns of QM/

***Table 1.*** Relative Free Energies and Internuclear Distances at Various States of RNA Self-Cleavage Catalysis in Hammerhead Ribozymes[a]

| | react | $TS_1$ | int | $TS_2$ | prod |
|---|---|---|---|---|---|
| Nu−P | 3.50(04) | 1.76(05) | 1.66(03) | 1.67(03) | 1.68(03) |
| P−Lea | 1.65(03) | 2.11(05) | 4.51(04) | 4.24(48) | 3.63(23) |
| gA−H | 0.96(00) | 0.96(00) | 0.96(00) | 1.78(04) | 3.75(04) |
| H−Lea | 2.57(51) | 4.07(47) | 4.13(73) | 1.04(03) | 1.00(03) |
| $Mg^{2+}$−Lea | 3.99(18) | 3.61(17) | 2.02(05) | 2.83(86) | 4.48(05) |
| $Mg^{2+}$−gA | 4.56(18) | 4.03(18) | 4.33(06) | 3.38(86) | 2.03(05) |
| $\Delta G$ | 0.0(4) | 24.4(6) | −6.7(3) | 13.7(7) | −13.6(9) |

[a] Free energies ($\Delta G$) are in kcal/mol, which were extracted from 1D PMF profiles along the minimum free-energy path through the 2D profiles. Average distances (X−Y) are in Å. Standard deviations are listed in parentheses divided by the decimal precision of the average values. The abbreviations "react", "TS", "int", and "prod" signify reactant, transition, intermediate, and product states, respectively, and for the distance metrics, "Nu", "Lea", "gA", and "H" refer to the O2′ nucleophile, O5′ leaving group, general acid residues G8:O2′, and H2′, respectively.

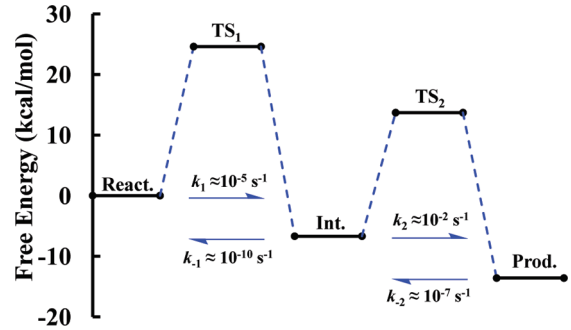



***Figure 1.*** Schematic diagram for relative free-energy barriers and the corresponding reaction rate constants.

* Corresponding author e-mail: york@biomaps.rutgers.edu.
[†] University of Minnesota.
[‡] Rutgers University.

10.1021/ct100467t 
Published on Web 12/16/2010

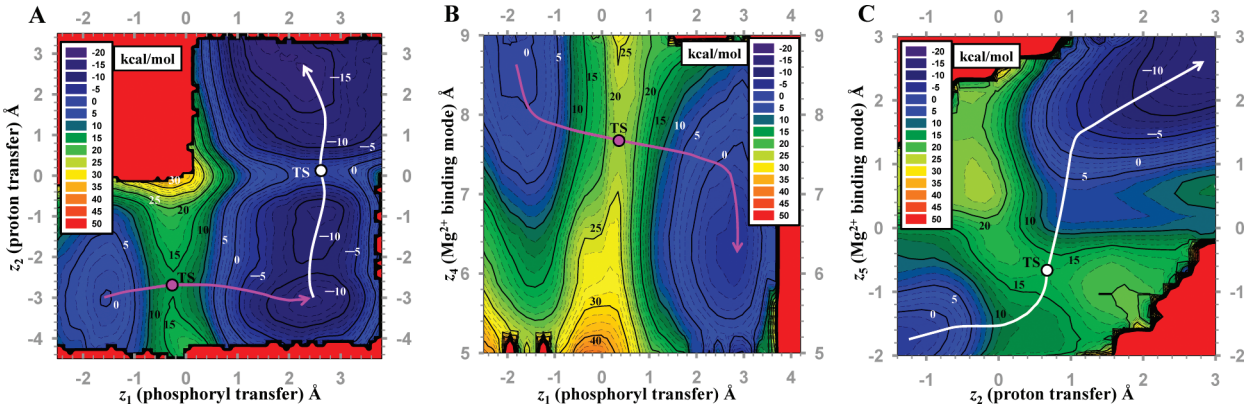



***Figure 2.*** (A) Selected 2D surface, harmonically restrained along the course-grained metal ion binding coordinate at $d$ ($Mg^{2+}$, G8:O2′) = 2.5 Å, where $z_1 = d$(O5′,P) − $d$(P,O2′), $z_2 = d$(G8:O2′,H) − $d$(H,O5′). (B) 2D PMF for the $Mg^{2+}$ binding mode in the phosphoryl transfer step, where $z_4 = d$($Mg^{2+}$, O5′) + $d$($Mg^{2+}$, G8:O2′). (C) 2D PMF for the $Mg^{2+}$ binding mode in the general acid step, where $z_5 = d$($Mg^{2+}$, O5′) − $d$($Mg^{2+}$, G8:O2′). $d(x,y)$ denotes distance between $x$ and $y$. TS is the acronym of transition state.

MM simulation. Simulations were based on the extended HHR solvent structure (PDB: 2OEU)[4] solvated by over 10 000 water molecules. Active site residues (G8, A9, C1.1, C17, and an $Mg^{2+}$ ion with coordinated water molecules shown in Figure 3) were treated quantum mechanically (81 atoms total) using the AM1/d-PhoT quantum model[9] with the AM1/d model for $Mg^{2+}$.[10] We used the all-atom AMBER parmbsc0 force field,[11] to describe the HHR outside of the active site, along with the TIP4P-Ewald water model[12] and the consistent set of monovalent ion parameters.[13] Multidimensional PMF profiles were generated along reaction coordinates corresponding to phosphoryl transfer, proton transfer from the general acid to the leaving group, and the divalent metal ion binding mode. Complete details are given in the Supporting Information.

*Phosphoryl transfer and general acid steps are stepwise, and sensitive to the $Mg^{2+}$ binding mode.* Our initial attempts to study the chemical steps of the HHR reaction from 2D PMF profiles using phosphoryl transfer and proton transfer reaction coordinates, but not considering a reaction coordinate associated with $Mg^{2+}$ ion binding mode, led to free energy barriers that were unexpectedly high (∼37 kcal/mol) compared to an estimated barrier of ∼20 kcal/mol derived from the experimental rate of one turnover per minute in HHR catalysis.[14] We extended the calculations so as to include a 3D-PMF profile with a course-grained reaction coordinate associated with the $Mg^{2+}$ binding mode and confirmed the sensitivity of the barriers to the $Mg^{2+}$ ion position along the reaction coordinate. A common feature of the reaction mechanism derived from the 3D profile was that the phosphoryl transfer and general acid steps were stepwise (e.g., Figure 2a), allowing these steps to be decoupled. Both the phosphoryl transfer and general acid steps of the reaction were coupled with the $Mg^{2+}$ binding mode, and hence separate 2D profiles were generated for each step with a reaction coordinate corresponding to the $Mg^{2+}$ binding mode as a second dimension. Table 1 summarizes key average geometrical parameters, and free energy values for stationary points along the reaction. The corresponding reaction rate constants are depicted in Figure 1.

*Phosphoryl transfer is rate-limiting and facilitated by electrostatic stabilization by $Mg^{2+}$.* The phosphoryl transfer step is rate-controlling, having a free energy barrier of approximately

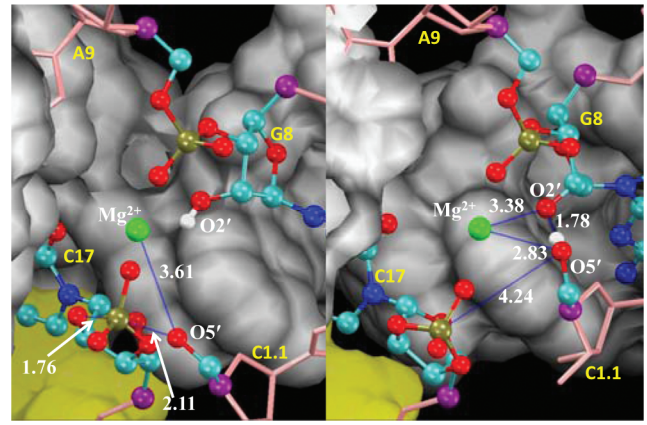



***Figure 3.*** Snapshots of the active site at the transition states for phosphoryl transfer (left) and general acid catalysis (right) with average distances labeled.

24.4 kcal/mol. The position of the $Mg^{2+}$ ion follows the negative charge along the phosphoryl transfer reaction coordinate in order to provide electrostatic stabilization. The change in the $Mg^{2+}$ position is continuous and monotonic throughout the phosphoryl transfer step (Figure 2b) and is most pronounced in the initial and final stages when the nucleophile and leaving group have the greatest negative charge. The transition state is late (Figure 3), having a P−O5′ distance of 2.11 Å. As the P−O5′ bond breaks, the $Mg^{2+}$ ion forms an innersphere coordination, leading to a $Mg^{2+}$-bound O5′ alkoxide intermediate.

*General acid catalysis is concerted with changes in the $Mg^{2+}$ binding mode*. The general acid step considered here assumes that the 2′OH of G8 acts as a general acid catalyst to transfer a proton to the O5′ leaving group. An examination of Figure 2c indicates that proton transfer occurs after formation of the $Mg^{2+}$-coordinated cleaved intermediate and is concerted with changes in $Mg^{2+}$ binding mode. Unlike the phosphoryl transfer step, participation of the $Mg^{2+}$ along the reaction coordinate is most pronounced not at the end points of the step but near the midpoint where the proton transfer occurs.

The free energy barrier for the transition state of the general acid step (Figure 3) is 13.7 kcal/mol with respect to the activated precursor state. The intermediate is only 6.7 kcal/mol lower in

free energy than the activated precursor and has a 20.4 kcal/mol barrier to breakdown into the product state with a proton fully transferred to the O5′ leaving group.

*Relation with experiment*. The present work explores a specific mechanistic scenario, departing from the activated precursor state, that assumes the catalytic metal ion is in a position bridging the A9 and scissile phosphates, and the 2′OH of G8 acts as a general acid catalyst. The former metal ion binding mode has not yet been observed experimentally but has been inferred from biochemical experiments on the minimal[15] and extended[16] HHR and predicted by molecular simulations.[5,17,18] The latter role of G8 is consistent from structural[4] and biochemical data.[6−8] There has been seminal work on the study of metal ion interactions for the minimal HHR[19,20] that provide insight into the ligand environment of the site bound metal. Time-resolved NMR experiments suggest that there is a dynamic equilibrium between energetically similar conformations in the minimal HHR that are sensitive to $Mg^{2+}$ binding,[21] and it has been suggested that the minimal and extended HHR may utilize a similar dynamic reaction mechanism for catalysis.[22] In the present study, we provide computational support for the plausibility of a cleavage mechanism where phosphoryl transfer and general acid catalysis are stepwise and the catalytic divalent metal ion plays an active role in the chemical steps of catalysis. It is the hope that this work, together with experimental work that probes the nature of metal ion interactions at the active site, will provide deeper insight into the underpinnings of chemical catalysis in the HHR.

**Acknowledgment**. The authors are grateful for financial support from the National Institutes of Health (GM084149 to D.Y.). Computational resources were provided by the Minnesota Supercomputing Institute (MSI) and by the NSF TeraGrid through the Texas Advanced Computing Center and National Institute for Computational Sciences under grant number TG-CHE100072. We thank Professor Victoria J. DeRose for useful comments on the manuscript.

**Supporting Information Available:** Additional computational details. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Strobel, S. A.; Cochrane, J. C. *Curr. Opin. Chem. Biol.* **2007**, *11*, 636–643.

(2) Scott, W. G. *Curr. Opin. Struct. Biol.* **2007**, *17*, 280–286.

(3) Leclerc, F. *Molecules* **2010**, *15*, 5389–5407.

(4) Martick, M.; Lee, T.-S.; York, D. M.; Scott, W. G. *Chem. Biol.* **2008**, *15*, 332–342.

(5) Lee, T.-S.; Silva Lopez, C.; Giambaşu, G. M.; Martick, M.; Scott, W. G.; York, D. M. *J. Am. Chem. Soc.* **2008**, *130*, 3053–3064.

(6) Blount, K. F.; Uhlenbeck, O. C. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 415–440.

(7) Nelson, J. A.; Uhlenbeck, O. C. *RNA* **2008**, *14*, 605–615.

(8) Thomas, J. M.; Perrin, D. M. *J. Am. Chem. Soc.* **2009**, *131*, 1135–1143.

(9) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486–504.

(10) Imhof, P.; Noé, F.; Fischer, S.; Smith, J. C. *J. Chem. Theory Comput.* **2006**, *2*, 1050–1056.

(11) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.

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

(13) Joung, I. S.; Cheatham, T. E., III *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(14) Scott, W. G. What can the New Hammerhead Ribozyme Structures Teach us About Design? In *RNA Technologies and Their Applications*; Erdmann, V., Barciszewski, J., Eds.; Springer-Verlag: Berlin, Heidelberg, 2010.

(15) Wang, S.; Karbstein, K.; Peracchi, A.; Beigelman, L.; Herschlag, D. *Biochemistry* **1999**, *38*, 14363–14378.

(16) Osborne, E. M.; Schaak, J. E.; Derose, V. J. *RNA* **2005**, *11*, 187–196.

(17) Lee, T.-S.; Silva-Lopez, C.; Martick, M.; Scott, W. G.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 325–327.

(18) Lee, T.-S.; Giambaşu, G. M.; Sosa, C. P.; Martick, M.; Scott, W. G.; York, D. M. *J. Mol. Biol.* **2009**, *388*, 195–206.

(19) Vogt, M.; Lahiri, S.; Hoogstraten, C. G.; Britt, D. R.; DeRose, V. J. *J. Am. Chem. Soc.* **2006**, *128*, 16764–16770.

(20) Osborne, E. M.; Ward, W. L.; Ruehle, M. Z.; DeRose, V. J. *Biochemistry* **2009**, *48*, 10654–10664.

(21) Fürtig, B.; Richter, C.; Schell, P.; Wenter, P.; Pitsch, S.; Schwalbe, H. *RNA Biol.* **2008**, *5*, 41–48.

(22) Nelson, J. A.; Uhlenbeck, O. C. *RNA* **2008**, *14*, 43–54.

CT100467T

JCTC Journal of Chemical Theory and Computation

# Information-Guided Noise Reduction in Forward−Backward Semiclassical Dynamics

Jonathan Chen and Nancy Makri*

*Department of Chemistry, University of Illinois, 601 S. Goodwin Avenue, Urbana, Illinois 61801, United States*

Received August 5, 2010

**Abstract:** Information-guided noise reduction (IGNoR) [*Chem. Phys. Lett.* **2004**, *400*, 446], a procedure for reducing the statistical error in Monte Carlo integration of oscillatory functions, is generalized to cases where both the prototype function and remaining integrand are complex-valued. The method is applied to the forward−backward semiclassical dynamics approximation of time correlation functions. Illustrative calculations of velocity autocorrelation functions in supercritical argon and liquid neon are presented.

## I. Introduction

The so-called sign problem presents a major challenge in the Monte Carlo (MC) integration of oscillatory functions.[1] In statistical and condensed matter physics, integrands of alternating sign occur primarily in equilibrium calculations of identical fermions (as a result of particle exchange to account for antisymmetry) and in quantum dynamical calculations (because of the phase interference in the time evolution operator). Currently, the severity of the sign problem restricts the application of quantum Monte Carlo and path integral Monte Carlo methods to high temperatures, short times, and/or small numbers of particles or necessitates the use of extraneous approximations.

Few attempts to overcome the sign problem have been reported. Blocking algorithms[2] minimize cancellation by grouping integrand points and performing the sum within each block with deterministic methods. By choosing the blocks judiciously, the variance of the block averages can be smaller than that of the original function, thus reducing cancellation in the Monte Carlo step. A different, fully Monte Carlo based approach is information-guided noise reduction[3] (IGNoR). This exploits the knowledge of the exact integral for a similar, prototypical highly oscillatory function. The Monte Carlo random walk samples the prototype function and the desired integrand, estimating the ratios of the positive and negative parts of the two functions. With that information, IGNoR replaces the raw Monte Carlo estimate of the negative part of the desired integral by a corrected estimate obtained from knowledge of the exact integral of the prototype function. If the original and prototype functions are very similar, the IGNoR procedure will by construction generate excellent statistics.

In addition to these quite general algorithms, there are several strategies for reducing the severity of the sign problem by smoothing the integrand. The majority of these approaches are by nature approximate methods, but a few can be classified as numerically exact. In the present paper, we focus on the IGNoR methodology, which is a general and strictly numerical scheme for improving the Monte Carlo statistics within a specific formulation, i.e., without altering the integrand.

Clearly, the choice of the prototype function in IGNoR is critical to the success of the algorithm. Thus, the most critical step in the application of the algorithm is the identification of a function that is as similar as possible to the given integrand but whose integral is known exactly or at least can be obtained numerically to high accuracy. In many quantum mechanical methods, the prototype function must be chosen as a complex-valued part of the integrand. Thus, we generalize the original IGNoR methodology to the case of complex functions, pointing out that application of the noise reduction procedure to separate (ungrouped) terms is highly beneficial.

In section II, we review the IGNoR methodology and extend it to complex-valued integrands. Section III shows how the IGNoR correction may be applied to the calculation of time correlation functions within the forward−backward semiclassical dynamics (FBSD) approximation. The same

* Corresponding author e-mail: nancy@makri.scs.uiuc.edu.

10.1021/ct1004356 
Published on Web 12/13/2010

section presents applications to the time correlation function of liquids, which demonstrates the dramatic decrease in statistical error attainable with the IGNoR procedure. Finally, some concluding remarks are given in section IV.

## II. Information-Guided Noise Reduction for Complex Functions

**a. IGNoR for Real Functions.** The goal is to evaluate the (definite) integral $J$ of an oscillatory function using a Monte Carlo procedure.[4] We start by expressing the desired integral in the form

$$J = \int f(x)\, g(x)\, \mathrm{d}x \tag{2.1}$$

where $f(x)$ is a reference function that is highly oscillatory, $g(x)$ is slowly varying, and the value of the integral

$$I = \int f(x)\, \mathrm{d}x \tag{2.2}$$

is known in advance. IGNoR exploits the helpful information in eq 2.2 to decrease the statistical error of the raw Monte Carlo estimate of eq 2.1.

First, $I$ and $J$ are split into their "signed" parts:

$$I_{\pm} = \int f(x)\, h(\pm f(x))\, \mathrm{d}x \tag{2.3}$$

$$J_{\pm} = \int f(x)\, g(x)\, h(\pm f(x))\, \mathrm{d}x \tag{2.4}$$

where

$$h(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases} \tag{2.5}$$

denotes the Heaviside step function. Next, the Monte Carlo random walk with a convenient sampling function is performed to obtain estimates $\langle I_{\pm}\rangle$ and $\langle J_{\pm}\rangle$ of the integrals 2.3 and 2.4. In the raw Monte Carlo procedure, the estimate of the desired integral is obtained by the simple addition

$$\langle J\rangle = \langle J_+\rangle + \langle J_-\rangle \tag{2.6}$$

The sign problem occurs because the positive and negative parts are comparable in absolute value, and their sum typically is smaller than the statistical error in their estimates. To improve the situation, IGNoR replaces $\langle J_-\rangle$ by the corrected value given by

$$\tilde{J}_- = \frac{\langle J_-\rangle}{\langle I_-\rangle}(I - \langle I_+\rangle) \tag{2.7}$$

Thus, the IGNoR-corrected estimate of the integral of interest is[3]

$$\langle J_+\rangle + \tilde{J}_- = \langle J_+\rangle + \frac{\langle J_-\rangle}{\langle I_-\rangle}(I - \langle I_+\rangle) \tag{2.8}$$

Because $\langle I_{\pm}\rangle$ and $\langle J_{\pm}\rangle$ are computed from the same random walk, these estimates are correlated; thus the error in the ratio $\langle J_-\rangle/\langle I_-\rangle$ generally is small. In fact, in the special case where $g(x) = a$ is a real constant, because the estimates $\langle I_{\pm}\rangle$ and $\langle J_{\pm}\rangle$ are obtained from the same random walk, $\langle J_-\rangle/\langle I_-\rangle = a$ exactly, and thus

$$\langle J_+\rangle + \tilde{J}_- = a\langle I_+\rangle + aI - a\langle I_+\rangle = aI = J \tag{2.9}$$

i.e., the IGNoR estimate 2.8 is exact. For nonconstant but slowly varying $g(x)$, eq 2.8 should be more accurate than the brute Monte Carlo estimate $\langle J_+\rangle + \langle J_-\rangle$.

We note that the procedure requires knowledge of the normalization integral of the MC sampling function. The latter must be available either exactly, or at least with much higher precision (smaller statistical uncertainty) than $\langle I_{\pm}\rangle$. Recovery of the exact integral value of 2.9 in the case of constant $g$ holds only if the exact value of the MC normalization integral is known.

Finally, we note that eq 2.9 can be symmetrized,[5] leading to the prescription

$$\frac{1}{2}[\langle J_+\rangle + \tilde{J}_- + \langle J_-\rangle + \tilde{J}_+] \tag{2.10}$$

taking care to remember that partial error cancellation is achieved through the *pairings* $\langle J_+\rangle + \tilde{J}_-$ and $\langle J_-\rangle + \tilde{J}_+$, as evident through eq 2.9. The symmetrized IGNoR prescription may offer some advantages, although unsymmetrized and symmetrized formulas produced very similar results in the calculations reported in section III.

**b. IGNoR for Complex Functions.** To extend IGNoR to cases where both functions are complex-valued, one needs to separate the integrand into appropriate real and imaginary parts. One may split the product $fg$ into its real and imaginary components, resulting in two integrals to be calculated by IGNoR. A second choice is to apply the IGNoR correction to each of the four terms in the product $fg$. An earlier application of IGNoR[6] used the first partitioning, which is not optimal, because the relevant functions are not necessarily strongly correlated. For example, in the special case where $g$ is a complex-valued constant, the IGNoR based on splitting the product $fg$ into real and imaginary parts does not lead to the exact result. Below, we describe the second partitioning, which is *designed* to have zero statistical error when $g$ is a complex-valued constant and thus should be more accurate when $g(x)$ is a smooth complex-valued function.

We begin by separating the desired integral into four components arising from the real and imaginary parts of each of the two functions:

$$J = J^{\mathrm{rr}} - J^{\mathrm{ii}} + iJ^{\mathrm{ri}} + iJ^{\mathrm{ir}} \tag{2.11}$$

where

$$J^{\mathrm{rr}} = \int \mathrm{Re}\, f(x)\, \mathrm{Re}\, g(x)\, \mathrm{d}x,\ J^{\mathrm{ri}} = \int \mathrm{Re}\, f(x)\, \mathrm{Im}\, g(x)\, \mathrm{d}x,$$
$$J^{\mathrm{ir}} = \int \mathrm{Im}\, f(x)\, \mathrm{Re}\, g(x)\, \mathrm{d}x,\ J^{\mathrm{ii}} = \int \mathrm{Im}\, f(x)\, \mathrm{Im}\, g(x)\, \mathrm{d}x \tag{2.12}$$

We also write $I = I^{\mathrm{r}} + iI^{\mathrm{i}}$ and define the integrals of the positive and negative components of each integral:

$$I^{\mathrm{r}}_{\pm} = \int \mathrm{Re}\, f(x)\, h(\pm \mathrm{Re}\, f(x))\, \mathrm{d}x,$$
$$I^{\mathrm{i}}_{\pm} = \int \mathrm{Im}\, f(x)\, h(\pm \mathrm{Im}\ f(x))\, \mathrm{d}x \tag{2.13}$$

$$\begin{aligned} J_{\pm}^{\mathrm{rr}} &= \int \mathrm{Re}\, f(x)\, \mathrm{Re}\, g(x)\, h(\pm \mathrm{Re}\, f(x))\, \mathrm{d}x, \\ J_{\pm}^{\mathrm{ri}} &= \int \mathrm{Re}\, f(x)\, \mathrm{Im}\, g(x)\, h(\pm \mathrm{Re}\, f(x))\, \mathrm{d}x, \\ J_{\pm}^{\mathrm{ir}} &= \int \mathrm{Im}\, f(x)\, \mathrm{Re}\, g(x)\, h(\pm \mathrm{Im}\, f(x))\, \mathrm{d}x, \\ J_{\pm}^{\mathrm{ii}} &= \int \mathrm{Im}\, f(x)\, \mathrm{Im}\, g(x)\, h(\pm \mathrm{Im}\, f(x))\, \mathrm{d}x \end{aligned} \tag{2.14}$$

Now, we apply the original IGNoR procedure to each component of $J$:

$$\begin{aligned} \tilde{J}_{-}^{\mathrm{rr}} &= \frac{\langle J_{-}^{\mathrm{rr}} \rangle}{\langle I_{-}^{\mathrm{r}} \rangle}(I^{\mathrm{r}} - \langle I_{+}^{\mathrm{r}} \rangle), \tilde{J}_{-}^{\mathrm{ri}} = \frac{\langle J_{-}^{\mathrm{ri}} \rangle}{\langle I_{-}^{\mathrm{r}} \rangle}(I^{\mathrm{r}} - \langle I_{+}^{\mathrm{r}} \rangle), \\ \tilde{J}_{-}^{\mathrm{ir}} &= \frac{\langle J_{-}^{\mathrm{ir}} \rangle}{\langle I_{-}^{\mathrm{i}} \rangle}(I^{\mathrm{i}} - \langle I_{+}^{\mathrm{i}} \rangle), \tilde{J}_{-}^{\mathrm{ii}} = \frac{\langle J_{-}^{\mathrm{ii}} \rangle}{\langle I_{-}^{\mathrm{i}} \rangle}(I^{\mathrm{i}} - \langle I_{+}^{\mathrm{i}} \rangle) \end{aligned} \tag{2.15}$$

Since the IGNoR prescription is now performed on four real-valued integrands, eq 2.15 should yield the exact value of the desired integral in the case where $g$ is a complex-valued constant. The numerical calculations reported in section III follow this procedure.

## III. Application: Forward−Backward Semiclassical Dynamics of Liquids

Various simulation methods developed in the past decade are based on semiclassical ideas.[7] Semiclassical methods[8] are attractive because they employ classical trajectories to capture dynamical effects. Still, the highly oscillatory semiclassical phase leads to a severe sign problem, and fully semiclassical calculations in many-particle systems remain impractical. Certain semiclassical approximations to Heisenberg operators, which contain forward and reverse time evolution steps, take advantage of the proximity (in the stationary phase limit) of forward and backward trajectories to eliminate the oscillatory semiclassical phase altogether.[9] In the particular case of time correlation functions, such ideas give rise to linearized semiclassical[10,11] (LSC) and forward−backward[12,13] semiclassical dynamics (FBSD) approximations. However, even with this extremely important stabilization, the resulting expressions converge much slower than fully classical calculations. This is so because of the presence of oscillatory components through coherent state factors (in the case of FBSD) or the Wigner transformation (in the case of LSC approximations). Further, more accurate (and costly) semiclassical treatments have been formulated, which reintroduce a portion of the oscillatory semiclassical phase, thereby capturing some coherence features at the expense of numerical stability.[14] All of these situations invite the application of IGNoR to improve convergence. In this section, we illustrate the enhancement obtained by implementing the IGNoR methodology on FBSD calculations of velocity correlation functions in neat liquids.

Our present focus is on time correlation functions at a finite temperature, specifically the correlation function for the inner product of two vector operators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$

$$C_{\mathbf{A}\cdot\mathbf{B}}(t) = \frac{1}{Z} Tr(\mathrm{e}^{-\beta\hat{H}} \hat{\mathbf{A}}\, \mathrm{e}^{i\hat{H}t/\hbar} \cdot \hat{\mathbf{B}}\, \mathrm{e}^{-i\hat{H}t/\hbar}) \tag{3.1}$$

Here, $\hat{H}$ is the Hamiltonian that describes the $n$-particle system, $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are vector operators, $\beta = 1/k_{\mathrm{B}}T$ is the reciprocal temperature, and $Z = Tr\, \mathrm{e}^{-\beta\hat{H}}$ is the canonical partition function. The three-dimensional Cartesian position and momentum vectors for the $j$th atom are denoted, respectively, as $\mathbf{r}_j = (r_{j1}, r_{j2}, r_{j3})$ and $\mathbf{p}_j = (p_{j1}, p_{j2}, p_{j3})$. For notational convenience, the coordinates and momenta of all of the particles are collected in the $3n$-dimensional vectors $\mathbf{Q}$ and $\mathbf{P}$, respectively. The Hamiltonian is written as

$$\hat{H} = \hat{T} + V(\hat{\mathbf{Q}}) \tag{3.2}$$

where $\hat{T}$ is the operator for the total kinetic energy of the system.

$$\hat{T} = \sum_{j=1}^{n} \hat{T}_j = \sum_{j=1}^{n} \frac{|\hat{\mathbf{p}}_j|^2}{2m_j} \tag{3.3}$$

Exponentiation of the operator $\hat{B}$ via a derivative identity and application of the forward−backward semiclassical idea in a coherent state representation[15] gives rise to the following FBSD approximation of eq 3.1:[13]

$$\begin{aligned} C_{\mathbf{A}\cdot\mathbf{B}}(t) \;=\; & (2\pi\hbar)^{-3n} Z^{-1} \int \mathrm{d}\mathbf{Q}^{(0)} \int \mathrm{d}\mathbf{P}^{(0)} \left(1 + \frac{3}{2}n\right) \langle G(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}) | \\ & | \mathrm{e}^{-\beta\hat{H}} \hat{\mathbf{A}} | G(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}) \rangle \cdot \mathbf{B}(\mathbf{Q}(t), \mathbf{P}(t)) \\ & - 2(2\pi\hbar)^{-3n} Z^{-1} \int \mathrm{d}\mathbf{Q}^{(0)} \int \mathrm{d}\mathbf{Q}^{(0)} \langle G(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}) | (\hat{\mathbf{Q}} - \mathbf{Q}^{(0)})\, \mathrm{e}^{-\beta\hat{H}} \\ & [\hat{\mathbf{A}} \cdot \mathbf{B}(\mathbf{Q}(t), \mathbf{P}(t))] \cdot \Gamma \cdot (\hat{\mathbf{Q}} - \mathbf{Q}^{(0)}) | G(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}) \rangle \end{aligned} \tag{3.4}$$

Here, $\mathbf{Q}^{(0)}$ and $\mathbf{P}^{(0)}$ are the initial phase space variables for classical trajectories that evolve according to the classical equations of motion corresponding to the product of three exponential operators, $\mathbf{Q}(t)$ and $\mathbf{P}(t)$ are the phase space variables at time $t$, and $\mathbf{B}(\mathbf{Q},\mathbf{P})$ is the classical analogue of the operator $\hat{\mathbf{B}}$. The ket in eq 3.4 represents a $3n$-dimensional coherent state,[16] i.e., a product of $n$ three-dimensional coherent states for each particle, and is described in the coordinate representation by the wave function

$$\begin{aligned} \langle \mathbf{Q} | G(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}) \rangle \;&=\; \prod_{j=1}^{n} \langle \mathbf{r}_j | g(\mathbf{r}_j^{(0)}, \mathbf{p}_j^{(0)}) \rangle \\ &= \left(\frac{2}{\pi}\right)^{3n/4} (\det \boldsymbol{\Gamma})^{1/4} \times \\ &\quad \exp[-(\mathbf{Q} - \mathbf{Q}^{(0)})\, \boldsymbol{\Gamma}\, (\mathbf{Q} - \mathbf{Q}^{(0)}) + \\ &\quad \frac{i}{\hbar} \mathbf{P}^{(0)} (\mathbf{Q} - \mathbf{Q}^{(0)}) \end{aligned} \tag{3.5}$$

where $\boldsymbol{\Gamma}$ is a $3n \times 3n$ matrix. Throughout the rest of the paper, we choose a diagonal form with elements $\gamma_j$ representing the width parameters of the coherent state for the $j$th particle. Typically, these values are chosen to match the characteristic frequencies of the system.

The next step to consider is the evaluation of the coherent state transform of the operators, both of which involve the Boltzmann operator. A fully quantum treatment of this time-independent part is essential in capturing the effects of zero-point energy and reproducing the imaginary part of a correlation function. An accurate evaluation of these matrix elements is possible[17−19] by using the discretized path

integral representation of the Boltzmann operator.[20] After some algebra, eq 3.4 becomes

$$C_{\mathbf{A}\cdot\mathbf{B}}(t) = (2\pi\hbar)^{-3n}Z^{-1} \times \int d\mathbf{Q}^{(0)} \int d\mathbf{P}^{(0)} \int d\mathbf{Q}^{(1)} ... \int d\mathbf{Q}^{(N)} \Theta(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) \Lambda_{\mathbf{A}\cdot\mathbf{B}}(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)};t) \quad (3.6)$$

In these equations, $\Delta\beta = \beta/N$,

$$\Theta(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) = \prod_{j=1}^{n}\left(\frac{2\gamma_j}{\pi}\right)^{3/2}\left(\frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}\right)^3 \left(\frac{m_j}{2\pi\hbar^2\Delta\beta}\right)^{3(N-1)/2} \times \exp\left\{-\sum_{j=1}^{n}\frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}\left(\gamma_j|\mathbf{r}_j^{(1)} - \mathbf{r}_j^{(0)}|^2 + \gamma_j|\mathbf{r}_j^{(N)} - \mathbf{r}_j^{(0)}|^2 + \frac{i}{\hbar}\mathbf{p}_j^{(0)}(\mathbf{r}_j^{(1)} - \mathbf{r}_j^{(N)}) + \frac{\Delta\beta}{2m_j}|\mathbf{p}_j^{(0)}|^2\right) - \sum_{j=1}^{n}\frac{m_j}{2\hbar^2\Delta\beta}\sum_{k=2}^{N}|\mathbf{r}_j^{(k)} - \mathbf{r}_j^{(k-1)}|^2 - \Delta\beta\sum_{k=1}^{N}V(\mathbf{Q}^{(k)})\right\} \quad (3.7)$$

is the integrand in the Trotter-discretized path integral representation of the coherent state transform of the Boltzmann operator, and the function $\Lambda_{\mathbf{A}\cdot\mathbf{B}}$ depends upon the form of the chosen operators. Recent work has focused on momentum autocorrelation functions, obtained with the choice $\hat{\mathbf{A}} = \hat{B} = \hat{\mathbf{p}}$. In that case, $\Lambda_{\mathbf{p}\cdot\mathbf{p}}$ is given by the expression[18]

$$\Lambda_{\mathbf{p}\cdot\mathbf{p}}(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)};t) = \left(1 + \frac{3}{2}n\right)\theta - 2\sum_{j=1}^{n}\sum_{\alpha=1}^{3}\gamma_j\varphi_{j\alpha}^{*}(r_{j\alpha}^{(0)}, p_{j\alpha}^{(0)}, r_{j\alpha}^{(1)})\left(-i\hbar\frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}p_{j\alpha}(t) + \theta\varphi_{j\alpha}(r_{j\alpha}^{(0)}, p_{j\alpha}^{(0)}, r_{j\alpha}^{(N)})\right) \quad (3.8)$$

where

$$\varphi_{j\alpha}(r_{j\alpha}^{(0)}, p_{j\alpha}^{(0)}, r_{j\alpha}^{(k)}) = \frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}\left(r_{j\alpha}^{(k)} - r_{j\alpha}^{(0)} + i\hbar\frac{\Delta\beta}{2m_j}p_{j\alpha}^{(0)}\right) \quad (3.9)$$

and

$$\theta = \sum_{j=1}^{n}\sum_{\alpha=1}^{3}\frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}[p_{j\alpha}^{(0)} + 2i\hbar\gamma_j(r_{j\alpha}^{(N)} - r_{j\alpha}^{(0)})]\, p_{j\alpha}(t) \quad (3.10)$$

Equation 3.6 is a quasiclassical expression. From eqs 3.8 and 3.10, one can see that all time dependence in the FBSD correlation function arises from the classical momentum values $p_{j\alpha}(t)$.

The multidimensional integral appearing in eq 3.6 is performed by Monte Carlo (or, alternatively, a molecular dynamics procedure[21]). The Monte Carlo procedure commonly uses the modulus of the absolute value of the exponential part as the sampling function

$$\rho(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) = |\Theta(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)})| \quad (3.11)$$

In the single-bead ($N = 1$) case, the integral of this sampling function is proportional to the partition function, thus the normalization integral cancels out. However, this is not the case when $N > 1$, and one must perform a separate Monte Carlo calculation to normalize the results. This can be done efficiently using a straightforward procedure.[18]

Even in the single-bead limit, the integrand of the FBSD expression contains negative components. In addition, a mildly oscillatory factor

$$\prod_{j=1}^{n}\exp\left[-\frac{i}{\hbar}\frac{m_j}{m_j + \hbar^2\Delta\beta\gamma_j}\mathbf{p}_j^{(0)}(\mathbf{r}_j^{(1)} - \mathbf{r}_j^{(N)})\right] \quad (3.12)$$

is present when $N > 1$, which arises from the coherent state phase. Because these factors introduce phase cancellation, whose magnitude increases exponentially with the number of particles, the use of IGNoR should be beneficial.

We have found the IGNoR procedure for complex functions ideally suited to the present situation with the choice

$$f(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) = (2\pi\hbar)^{-3n}Z^{-1}\Theta(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) \Lambda_{\mathbf{p}\cdot\mathbf{p}}(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)};0) \quad (3.13)$$

$$g(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)}) = \frac{\Lambda_{\mathbf{p}\cdot\mathbf{p}}(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)};t)}{\Lambda_{\mathbf{p}\cdot\mathbf{p}}(\mathbf{Q}^{(0)}, \mathbf{P}^{(0)}, \mathbf{Q}^{(1)}, ..., \mathbf{Q}^{(N)};0)} \quad (3.14)$$

Because eq 3.13 is the integrand of the $t = 0$ expression, its integral is proportional to the equilibrium value of the kinetic energy

$$I = Z^{-1}Tr(e^{-\beta\hat{H}}\,\hat{\mathbf{p}}\cdot\hat{\mathbf{p}}) \quad (3.15)$$

and can be obtained by a high-precision path integral Monte Carlo calculation (which does not suffer from the dynamical sign problem). Further, the imaginary part of the integral is identically equal to zero. Here, we report the results of calculations on supercritical argon and liquid neon.

Figure 1 shows single-bead ($N = 1$) FBSD results for supercritical argon. The calculations were performed at 183 K and a density of 1.15 g cm$^{-3}$ upon a cubic unit cell containing 108 atoms under periodic boundary conditions. All interactions were truncated at half of the unit cell dimensions and are described by a Lennard-Jones potential with $\varepsilon = 85$ cm$^{-1}$ and $\sigma = 3.4$ Å, where $r$ represents the interatomic separation.[22] Raw and IGNoR-corrected Monte Carlo results from calculations with only 10 000 trajectories are compared against earlier results using molecular dynamics sampling with 1 million trajectories. At $t = 0$, the partitioning described in eqs 3.13 and 3.14 results in $g = 1$, which guarantees, by construction, zero IGNoR error. However, Figure 1 shows that the statistical error of the IGNoR calculations remains small as the time increases. Although the integrand is not rapidly oscillatory in the single bead limit, where the problematic phase factor 3.12 reduces to unity, application of the IGNoR procedure described in the previous section results in considerable shrinking of statistical error, yielding meaningful results with a 100-fold

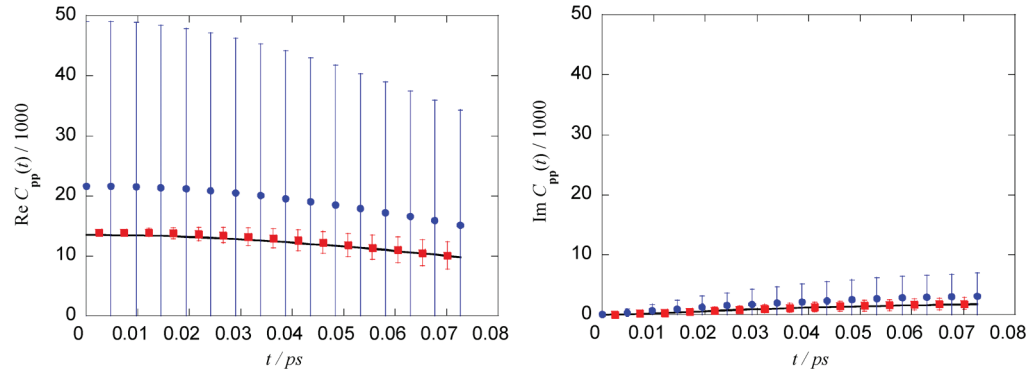



***Figure 1.*** Velocity autocorrelation function of supercritical argon from FBSD calculations in the single-bead discretization of the Boltzmann operator. Blue and red markers show raw and IGNoR-corrected Monte Carlo results (along with error bars), respectively, from calculations with 10 000 trajectories. The black line shows equivalent results using molecular dynamics sampling with 1 million trajectories.

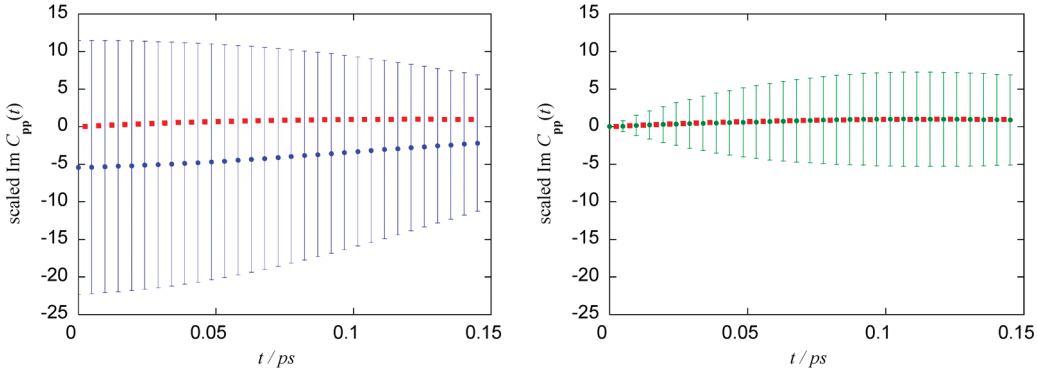



***Figure 2.*** Velocity autocorrelation function of liquid neon from FBSD calculations with $N = 2$ using 8 million trajectories. (a) Raw (blue) and IGNoR-corrected (red) Monte Carlo results. (b) Comparison of IGNoR corrected results from the application of error correction to each term of the real and imaginary parts, as prescribed by eq 2.15 (red), vs application to the entire real and imaginary part of the correlation function (green). It is seen that the current procedure, where the IGNoR processing is applied to each term separately, leads to much better statistics. The error bars of the results obtained from the IGNoR procedure given by eq 2.15 are smaller than the size of the markers.

reduction in the number of trajectories compared to the earlier raw calculation.

Similar calculations were performed on liquid neon with parameters chosen to match the calculations of Lawrence et al. ($T = 29.9$ K and a density of 0.03755 Å$^{-3}$). Interactions between neon atoms were described by a Lennard-Jones potential with $\varepsilon = 35.6$ K and $\sigma = 2.749$ Å. In order to avoid performing additional Monte Carlo calculations required for normalization, the IGNoR-corrected results were scaled arbitrarily to have the maximum of each curve equal to unity. This is possible because $I^{i} = 0$ for FBSD autocorrelation functions.

Figure 2 shows (scaled) raw Monte Carlo and IGNoR-corrected FBSD results with $N = 2$ obtained from a calculation with 8 million trajectories. As seen in Figure 2a, the error reduction accomplished by applying the IGNoR procedure is much more dramatic compared to the single-bead calculation in Figure 1. This is so because negative integrand areas arise not only from $\Lambda_{\mathbf{p}\cdot\mathbf{p}}$ but also from factor 3.12, which makes phase cancellation much more prevalent. Also, Figure 2b shows that the current application of IGNoR, eq 2.15, where the noise reduction is applied to *each* term in the real and imaginary components, results in much better statistics.

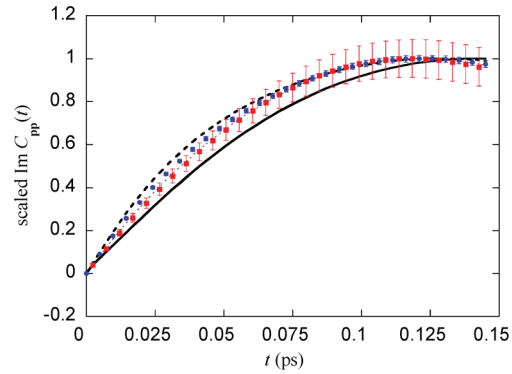



***Figure 3.*** Velocity autocorrelation function of liquid neon from IGNoR-corrected FBSD calculations with $N = 1-4$ (black dashed line, blue circles, black dashed line and red squares, respectively) using 8 million trajectories. For clarity, error bars are shown only for $N = 2$ and 4. The black line shows FBSD results with the PPP single-bead discretization of the Boltzmann operator. All curves have been scaled to the same maximum.

Finally, Figure 3 shows similar (scaled) IGNoR-corrected FBSD results for $N = 1-4$. As the number of path integral beads is increased, the Trotter-discretized FBSD results approach those obtained in the single-bead PPP approxima-

tion to the Boltzmann operator.[22] An increase in the number of beads beyond $N = 1$, which was previously prohibitive computationally with these particular parameters because of dramatic phase cancellation, becomes feasible with modest numbers of trajectories by incorporating the IGNoR correction in the FBSD methodology.

## IV. Concluding Remarks

We have presented an efficient extension of the IGNoR Monte Carlo methodology to complex-valued functions. The basic requirement of IGNoR is the identification of a function $f$ that incorporates as much of the integrand (including the oscillatory components) as possible and whose integral is known accurately. If the remaining factor $g$ of the integrand is sufficiently smooth, the IGNoR correction leads to a large reduction of statistical error compared to the raw Monte Carlo estimate. In the case of complex-valued functions, we find the application of IGNoR to each of the four terms arising from the real and imaginary parts of the product $fg$ most effective.

The application of IGNoR to FBSD correlation functions in neat liquids led to a dramatic reduction of statistical error. When the Boltzmann operator is discretized using multiple path integral beads ($N > 1$), the FBSD integrand is oscillatory and its variance grows rapidly as the number of particles is increased. For certain parameters, the raw Monte Carlo evaluation of the FBSD expression becomes prohibitively expensive. The application of IGNoR reduced the statistical error by several orders of magnitude, making convergence feasible with modest amounts of computational effort.

Because our goal in this paper was assessing the improvement attainable through IGNoR, our calculations employed the primitive Trotter discretization. In addition, we refrained from performing the separate Monte Carlo calculation required to evaluate the normalization integral of the sampling function, reporting un-normalized results. Future work should incorporate the correct normalization constant and employ the more efficient pair-product propagator,[23,24] which will lead to much faster convergence of FBSD correlation functions in neat fluids.

## References


(1) *Quantum Monte Carlo methods in condensed matter physics*; Suzuki, M., Ed.; World Scientific: Singapore, 1993.

(2) Mak, C. H.; Egger, R.; Gottschick, J. *Phys. Rev. Lett.* **1998**, *81*, 4533.

(3) Makri, N. *Chem. Phys. Lett.* **2004**, *400*, 446.

(4) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

(5) Jadhao, V. Private communication.

(6) Bukhman, E.; Makri, N. *J. Phys. Chem.* **2009**, *113*, 7183.

(7) Van Vleck, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1928**, *14*, 178.

(8) Miller, W. H. *Adv. Chem. Phys.* **1974**, *25*, 69.

(9) Makri, N.; Thompson, K. *Chem. Phys. Lett.* **1998**, *291*, 101.

(10) Wang, H.; Sun, X.; Miller, W. H. *J. Chem. Phys.* **1998**, *108*, 9726.

(11) Poulsen, J. A.; Nyman, G.; Rossky, P. J. *J. Chem. Phys.* **2003**, *119*, 12179.

(12) Sun, X.; Miller, W. H. *J. Chem. Phys.* **1999**, *110*, 6635.

(13) Shao, J.; Makri, N. *J. Phys. Chem. A* **1999**, *103*, 7753.

(14) Thoss, M.; Wang, H.; Miller, W. H. *J. Chem. Phys.* **2001**, *114*, 9220.

(15) Herman, M. F.; Kluk, E. *Chem. Phys.* **1984**, *91*, 27.

(16) Glauber, R. J. *Phys. Rev.* **1963**, *130*, 2529.

(17) Jezek, E.; Makri, N. *J. Phys. Chem.* **2001**, *105*, 2851.

(18) Makri, N. *J. Phys. Chem. B* **2002**, *106*, 8390.

(19) Yamamoto, T.; Wang, H.; Miller, W. H. *J. Chem. Phys.* **2002**, *116*, 7335.

(20) Feynman, R. P. *Statistical Mechanics*; Addison-Wesley: Redwood City, CA, 1972.

(21) Wright, N. J.; Makri, N. *J. Chem. Phys.* **2003**, *119*, 1634.

(22) Lawrence, C. P.; Nakayama, A.; Makri, N.; Skinner, J. L. *J. Chem. Phys.* **2004**, *120*, 6621.

(23) Ceperley, D. M. *Rev. Mod. Phys.* **1995**, *67*, 279.

(24) Nakayama, A.; Makri, N. *J. Chem. Phys.* **2003**, *119*, 8592.


CT1004356

JCTC Journal of Chemical Theory and Computation

# Convergent Partially Augmented Basis Sets for Post-Hartree−Fock Calculations of Molecular Properties and Reaction Barrier Heights

Ewa Papajak* and Donald G. Truhlar*

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

Received June 25, 2010

**Abstract:** We present sets of convergent, partially augmented basis set levels corresponding to subsets of the augmented "aug-cc-pV(*n*+d)Z" basis sets of Dunning and co-workers. We show that for many molecular properties a basis set fully augmented with diffuse functions is computationally expensive and almost always unnecessary. On the other hand, unaugmented cc-pV(*n*+d)Z basis sets are insufficient for many properties that require diffuse functions. Therefore, we propose using intermediate basis sets. We developed an efficient strategy for partial augmentation, and in this article, we test it and validate it. Sequentially deleting diffuse basis functions from the "aug" basis sets yields the "jul", "jun", "may", "apr", etc. basis sets. Tests of these basis sets for Møller−Plesset second-order perturbation theory (MP2) show the advantages of using these partially augmented basis sets and allow us to recommend which basis sets offer the best accuracy for a given number of basis functions for calculations on large systems. Similar truncations in the diffuse space can be performed for the aug-cc-pV*x*Z, aug-cc-pCV*x*Z, etc. basis sets.

## 1. Introduction

In quantum mechanical electronic structure calculations, the orbitals in configuration state functions may be represented as linear combinations of contracted functions, which in turn are linear combinations of spherical harmonics times radial functions with preoptimized exponential parameters. The radial functions can be Gaussian or Slater-type functions. Slater-type functions are more physical, but for ease of computation of the two-electron integrals, Slater-type functions are usually replaced by linear combinations of Gaussian-type functions.[1,2] The so-called ζ level reflects the degree of decontraction of the primitive Gaussian functions used to represent valence orbitals. For example, for carbon, valence double-ζ denotes two contracted functions designed to represent 2s orbitals and two contracted subshells to represent 2p orbitals, whereas valence triple-ζ denotes three, etc. Since the core is single-ζ quality in most modern basis sets, one often says "double-ζ" instead of "valence double-ζ", and so forth for triple, quadruple, etc. Double-, triple-, and quadruple-ζ are usually abbreviated DZ, TZ, and QZ, respectively. The contraction coefficients and exponential parameters for standard basis sets are available as lists called basis sets. Standard basis sets usually include polarization functions (e.g., *d* functions for carbon, *p* functions for hydrogen) and sometimes include diffuse functions.

The choice of the basis set for a given problem is critical, because it greatly affects the quality of the results as well as the cost of acquiring them. A basis set that is too large can make higher-level methods unaffordable for a given system or a given level unaffordable for larger systems. On the other hand, too small of a basis set can prevent taking full advantage of the potential accuracy of an otherwise very accurate electronic structure level. We have been especially interested in the requirements for diffuse basis functions as they play an important role in calculations of such commonly computed molecular properties as electron affinities, noncovalent interaction energies, and barrier heights for chemical reactions. Diffuse functions are characterized by very small exponential parameters, which allow the electrons to be

* Corresponding author e-mail: papajak@umn.edu (E.P.), truhlar@umn.edu (D.G.T.).

10.1021/ct1005533 
Published on Web 12/09/2010

***Table 1.*** Angular Momenta Included in the Diffuse Space in Tested Basis Sets

| basis set[a] | Li through Ca | H and He |
|---|---|---|
| aug-cc-pV(Q+d)Z | s p d f g | s p d f |
| jul-cc-pV(Q+d)Z | s p d f g | |
| jun-cc-pV(Q+d)Z | s p d f | |
| may-cc-pV(Q+d)Z | s p d | |
| apr-cc-pV(Q+d)Z[b] | s p | |
| cc-pV(Q+d)Z | | |
| aug-cc-pV(T+d)Z | s p d f | s p d |
| jul-cc-pV(T+d)Z | s p d f | |
| jun-cc-pV(T+d)Z | s p d | |
| may-cc-pV(T+d)Z[c] | s p | |
| cc-pV(T+d)Z | | |
| aug-cc-pV(D+d)Z | s p d | s p |
| jul-cc-pV(D+d)Z | s p d | |
| jun-cc-pV(D+d)Z[d] | s p | |
| cc-pV(D+d)Z | | |

[a] The same diffuse functions can be employed for month-cc-pV(*n*+*d*)Z or cc-pV(*n*+*d*)Z and for month-cc-pV*n*Z or cc-pV*n*Z. [b] Same as maug-cc-pV(Q+d)Z. [c] Same as maug-cc-pV(T+d)Z. [d] Same as maug-cc-pV(D+d)Z.

further away from the nuclei, so they are crucial for many systems and properties involving anions, transition states, excited states, and polarizability. However, if an unnecessarily large set of diffuse functions is used, the calculations may become unfeasible or unnecessarily expensive.

Three approaches to supplying basis sets with diffuse functions have emerged and have been widely used. In basis sets developed by Pople, Schleyer, and co-workers[3,4] (sometimes called Pople-type basis sets), *s* and *p* diffuse functions are added on atoms heavier than He, which is indicated by "+" in the name of the basis set. A second "+" in the name indicates that in addition to the diffuse functions on the nonhydrogenic atoms, diffuse *s* functions are added on the hydrogen atoms. Correlation consistent basis sets by Dunning and co-workers (including cc-pV*n*Z[5−9] and cc-pV(*n*+d)Z[10]) are augmented with gradually increasing sets of diffuse functions for increasing decontraction of the valence space (aug-cc-pV*n*Z[5−9,11,12] and aug-cc-pV(*n*+d)Z[10]). The angular momentum quantum numbers included in the diffuse spaces of aug-cc-pV*n*Z and aug-cc-pV(*n*+d)Z basis sets for $n$ = D, T, and Q are listed in Table 1. The other rows of this table (other than "aug-") will be explained below. In a third approach, Jensen recommended including sets of diffuse functions (on all atoms) such that the number of diffuse functions increases with increasing $n$.[13] The Pople strategy is not convergent in the diffuse space, in that the number of diffuse functions is not systematically increased when the $\zeta$ level is increased, whereas the Dunning and Jensen approaches are convergent.

We will use the term "augmentation" to denote adding diffuse functions to a basis (whereas "extension" means raising the $\zeta$ level or adding polarization functions), and we will use "full augmentation" to denote adding a diffuse function for every angular momentum present in the unaugmented basis set. The aug- basis sets are fully augmented. It has been known for a long time that augmentation on hydrogen atoms is less important than augmentation on nonhydrogenic atoms (e.g., Schleyer and co-workers defined the "+" basis sets that omit augmentation on hydrogen atoms;[3] del Bene and Shavitt[14] showed that they could converge proton affinities and hydrogen bonding energies with respect to basis set without diffuse functions on H; and Lynch and one of the authors[15] showed that diffuse functions are not needed on H even in molecules where H has a partial negative charge), and the present article is mainly concerned with the level of augmentation on nonhydrogenic atoms. We have previously shown that the full augmentation of nonhydrogenic atoms in the basis sets defined by Dunning and co-workers often leads to unnecessary expense in calculations by density functional theory (DFT).[16,17] In particular, we showed that the nonconvergent "+" approach is often sufficient for DFT calculations, and a basis set obtained by deleting all diffuse functions in aug- basis sets except the diffuse *s* and *p* functions on non-hydrogenic atoms has been called minimally augmented, denoted by the prefix maug.[16] However, we have also seen that wave function theory (WFT) calculations are more sensitive than DFT calculations to the saturation of the diffuse space in the nonhydrogenic basis space as well as to the size of the basis set in general.[16,17] This is because correlation energy in WFT converges very slowly with respect to the number of one-electron basis functions, and this slow convergence occurs because products of one-electron functions in a Slater-determinant poorly describe the cusps in the two-electron densities as the interelectronic distance approaches zero. The most slowly convergent part of the electron correlation energy is the part covered by second-order Møller−Plesset perturbation theory[18] (MP2). The reason for this is that although higher-order corrections are important, the contributions of the weakly coupled virtual orbitals at second order are quantitatively larger,[19,20] and it is important to understand how to include them efficiently. In practical work, in order to obtain as accurate results as one can afford using WFT for a given system, it is a common practice[21−26] to calculate the MP2 energy part at the complete basis set limit and to add higher-order corrections (e.g., the difference between an MP2 calculation and a calculation by the coupled cluster method with single and double excitations and a quasiperturbative treatment of connected triple excitations[27] CCSD(T)) calculated with a smaller basis set. Therefore, achieving the MP2 CBS limit is of a great practical interest.

One popular way to determine the MP2 CBS limit is extrapolation.[28−30] Recently, Møller−Plesset perturbation calculations employing basis functions that depend explicitly on electron−electron distances (MP2-R12 or MP2-F12[31−44]) have provided a powerful, alternative way to approach the MP2 basis set limit in a very efficient way, by explicitly improving the description of the cusp. MP2-F12 is very rapidly convergent with respect to the size of the one-electron basis set. In some key studies, the rapid convergence of the MP2-F12 method has often been established on the basis of heats of formation,[45] absolute correlation energy,[48] and energies of reaction,[48] but we note that, for neutral molecules and cations, heats of formation and energies of reaction are typically insensitive to the inclusion of the diffuse basis functions; thus a more recent study by Werner et al.[46] that examined not only reaction energies and atomization energies but also electron affinities, ionization potentials, equilibrium

structures, vibrational frequencies, and intermolecular interaction energies and a study by Kjaergaard et al.[47] on hydrogen bonded systems are more relevant to the question of how many diffuse functions one should use for a diverse set of molecular properties (Werner et al. also cite earlier diverse benchmarking studies). Recently, basis sets have been prepared specifically for use in F12 calculations.[48] Those basis sets are specifically limited to minimal augmentation, but the reader was informed that "the inclusion of just *s* and *p* diffuse functions may not be sufficient" in all cases, and "further extension of the higher angular momentum functions might then be considered." Werner et al.[46] employed these minimally augmented basis sets and fully augmented ones in their F12 benchmarking but did not consider intermediate augmentation. In the present article, we will consider this; in particular, we systematically explore various levels of partial augmentation in both MP2 and MP2-F12 calculations with databases for atomization energies, barrier heights, hydrogen bond energies, ionization potentials, and electron affinities.

It is an oversimplification to assume that "the more diffuse functions, the better." The size of the diffuse space of a basis set is just one of the parameters of a basis set, and it must be considered in conjunction with the space spanned by primitive valence functions, the level of contraction, and the number of polarization functions. Full augmentation with many diffuse functions that make only a small difference in the property to be calculated increases the size (and, one hopes, the accuracy) of the basis set a given $\zeta$ level, but the gain may be small relative to other ways to increase the accuracy such as increasing the $\zeta$ level. In practice, when attempting to improve a basis set, one should ask which aspect of the basis set is most limiting at any given level, and then one should improve that specific part of the basis set first. We have found that this is not always the way calculations are done. Very often, basis sets are fully augmented or fully augmented on nonhydrogenic atoms while staying at given $\zeta$ level when it would be more efficient to do only a partial augmentation on nonhydrogenic atoms and spend the saved resources by going to a higher $\zeta$ level or adding more polarization functions.

The questions we are attempting to answer are the following:

• Which molecular properties are most sensitive to the level of augmentation of the diffuse space? What level of augmentation is advisible for these properties?

• At what point is saturation of the diffuse space reached for key properties such as barrier heights, electron affinities, ionization potentials, noncovalent interaction energies, atomization energies, and bond energies?

• What is the right order of steps in improving the basis set for a given problem? At a given $\zeta$ level and for a given property, is it more beneficial to add more diffuse functions or to attempt to go to the higher $\zeta$ sooner with a minimal or intermediate number of diffuse functions?

A variety of *ad hoc* partially augmented basis sets have been used for specific calculations in various publications. Here, we put forward a systematic partial augmentation scheme and test it carefully for MP2 and MP2-F12 calculations. The naming convention for the systematically partially augmented basis sets is based on the months and involves successively truncating the "aug" basis sets of Dunning and co-workers to well-defined levels called "jul," "jun," "may," etc.

A second objective of the present article is to allow us to compare the accuracy of MP2-F12 to that of MP2.

## 2. Definition of the New Basis Sets

In our calculations, we use the correlation consistent cc-pV($n$+d)Z basis sets of Dunning and co-workers with spherical harmonic *d*, *f*, and *g* subshells. Notice that a cc-pV($n$+d)Z basis set for an atom lighter than Al is defined to be the same as cc-pV($n$+d)Z.[10] Table 1 defines the diffuse spaces of the fully and partially augmented basis sets. As seen in the table, deleting all diffuse basis functions on hydrogen and helium atoms from "aug" basis sets yields the "jul" basis set, which has already been defined.[17] The "jun", "may", etc. basis sets are obtained by sequentially removing the diffuse subshells on the heavy atoms, where "heavy atoms" is used here as a synonym for atoms heavier than He. The new basis sets are systematically convergent in that, just as aug-cc-pV($n$+d)Z converges to a fully augmented complete-valence basis set as $n$ increases, jun-cc-pV($n$+d)Z converges to a fully heavy-atom-augmented complete-valence basis set as $n$ increases, and so does may- or apr-. As discussed in the Introduction, it has been known for a long time that augmentation on heavy atoms is less important than augmentation on hydrogen atoms.

## 3. Methods and Databases

All of the MP2 calculations for this paper (except when timing MP2-F12 vs MP2) were carried out using the *Gaussian 03*[49] and *Gaussian 09*[50] program packages. All of the MP2-F12 calculations were performed using the *MolPro 2009*[51] program.

In order to test the performance of the partially augmented basis sets, we have used previously optimized geometries of the species contained in the DBH24/08,[52,53] HB6/04,[54] EA13/3,[55,56] IP13/3,[55,56] and AE6[57] databases.

These databases contain, respectively, 24 diverse barrier heights, six hydrogen bond energies (four of which were used for the present tests), 13 electron affinities, 13 ionization potentials, and six atomization energies.

We carried out restricted single-point-energy MP2 and MP2-F12 calculations for both open- and closed-shell species. Results of these MP2 and MP2-F12 single-point calculations using basis sets ranging from aug-cc-pV($n$+d)Z through partially augmented basis sets to nondiffuse cc-pV($n$+d)Z were then compared to MP2-F12/aug-cc-pV(Q+d)Z values that should be close to the complete basis set limit and that serve as a reference. MP2-F12/aug-cc-pV(Q+d)Z data were generated for the following properties:

• 24 (forward and backward) barrier heights using quadratic configuration interaction with single and double excitations QCISD/MG3S geometries (listed in the DBH24/08 database[52,53]) for the species (reactants, transition states, and products) involved in the following 12 reactions:

$$H + N_2O \leftrightarrow OH + N_2 \quad (1)$$

$$H + HCl \leftrightarrow HCl + H \quad (2)$$

$$CH_3 + FCl \leftrightarrow CH_3F + Cl \quad (3)$$

$$Cl^- + CH_3Cl \leftrightarrow ClCH_3 + Cl^- \quad (4)$$

$$F^- + CH_3Cl \leftrightarrow FCH_3 + Cl^- \quad (5)$$

$$OH^- + CH_3F \leftrightarrow CH_3OH + F^- \quad (6)$$

$$H + N_2 \leftrightarrow HN_2 \quad (7)$$

$$H + C_2H_4 \leftrightarrow CH_3CH_2 \quad (8)$$

$$HCN \leftrightarrow HNC \quad (9)$$

$$CH_4 + OH \leftrightarrow CH_3 + H_2O \quad (10)$$

$$H + OH \leftrightarrow O + H_2 \quad (11)$$

$$H + H_2S \leftrightarrow H_2 + HS \quad (12)$$

• the hydrogen bonding energy calculations for the $NH_3-NH_3$, HF−HF, $H_2O-H_2O$, and $NH_3-H_2O$ dimers for the MC-QCISD/3 geometries; this subset of the HB6/04 database[54] will be called HB4

• the electron affinities for C, S, O, Si, P, Cl, OH, SH, PH, $PH_2$, $O_2$, $S_2$, and $Cl_2$ using QCISD/MG3S geometries listed in EA13/3 database[55,56]

• the ionization potentials of the same 13 atoms and molecules as for electron affinities for the QCISD/MG3S geometries listed in the IP13/3 database[55,56]

• the atomization energies per bond for $SiH_4$, SiO, $S_2$, propyne ($C_3H_4$), glyoxal ($C_2H_2O_2$), and cyclobutane ($C_4H_8$) were calculated using the QCISD/MG3S geometries in the AE6 database.[57]

In all of our calculations, the Born−Oppenheimer electronic energies including nuclear repulsion were considered; vibrational contributions were not included. Spin−orbit energies were added (when nonzero) as in previous work.[16,17]

For this article, all of the MP2-F12/aug-cc-pV(n+d)Z calculations were carried out using the aug-cc-pV*n*Z/JKFIT[58] and aug-cc-pV*n*Z/MP2FIT[59,60] density-fitting basis sets. For all of the calculations involving nondiffuse (cc-pV(*n*+d)Z) orbital basis sets, cc-pV*n*Z/JKFIT[58] and cc-pV*n*Z/MP2FIT[59,60] density fitting basis sets were used. In order to reduce the cost of the calculations and at the same time pose a bigger challenge for the partially augmented (month-cc-pV(*n*+d)Z) sets, we used the density fitting basis sets recommended for the cc-pV*n*Z basis sets. The density fitting basis sets might not be completely converged for all basis sets, but they should be close enough to convergence that their small extent of incompleteness does not affect our conclusions.

All of the MP2-F12 calculations use the 3C ansatz with orbital invariant amplitudes and were carried out with a full nonlinear fit of the geminal expansion.

## 4. Results

In order to allow the assessment of computational costs (computer time and storage) of the various basis sets, Table 2 lists the total numbers of contracted basis functions for all of the calculations carried out for this article. This provides representative relative numbers for typical applications. Table 3 summarizes relative computer timings normalized to the least expensive basis set used (cc-pV(D+d)Z). For the ease of estimating computational savings by using partially instead of fully augmented basis sets at the same ζ level, we also list timings normalized to the unaugmented basis set for the same *n*.

Basis sets in the first column of all tables are listed in order of decreasing size. In each case, the basis set listed above the unaugmented basis set is the minimally augmented (maug) basis set containing s and p functions on heavy atoms (which means heavier than He) and no diffuse functions on H and He atoms. Subminimal augmentation (diffuse s functions on heavy atoms) gives results at best slightly better than no augmentation and is not recommended. Therefore, in this article, except for unaugmented results, no results are presented for subminimal augmentation.

Both MP2 and MP2-F12 results are compared with MP2-F12/aug-cc-pVQZ values. Tables 4−8 list mean signed deviation (MSD) and mean unsigned deviation (MUD) from this reference. Except for atomization energies, we define MSD and MUD as

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N} d_i \quad (13)$$

$$\mathrm{MUD} = \frac{1}{N}\sum_{i=1}^{N} |d_i| \quad (14)$$

where *N* is the number of energetic data computed (24 for barrier heights, 4 for hydrogen bonding energies, 13 for electron affinities and ionization potentials), and $d_i$ is the deviation (in kcal/mol) of the *i*th value from the reference value.

In order to report MSD and MUE for atomization energies on per-bond basis, mean deviations, computed using eqs 13 and 14, were divided by the average number of bonds in the AE6 database as in eqs 15 and 16.

$$\mathrm{MSD} = \frac{1}{4.83N}\sum_{i=1}^{N} d_i \quad (15)$$

$$\mathrm{MUD} = \frac{1}{4.83N}\sum_{i=1}^{N} |d_i| \quad (16)$$

Root mean square deviation values (RMSD) for all of the properties are available in the Supporting Information. In Tables 4−8, we provide mean deviations of MP2 (Tables 4−6) and MP2-F12 (Tables 6−8) results from the reference values. We have also listed maximum unsigned deviations (MaxUD) for completeness.

Counterpoise corrections[61] are sometimes used to estimate corrections for basis set superposition error, but such corrections are problematic in terms of accuracy[62] and become increasingly complex as one considers systems with more components[63] and hence are often omitted. In Table 5, we present an analog of the triple-ζ results in Table 4 for

***Table 2.*** Number of Basis Functions Used in MP2 Calculations

| | $N_{bf}$[a] | | | | | sum | | |
|---|---|---|---|---|---|---|---|---|
| | DBH24/08 | HB4 | EA13/3 | IP13/3 | AE6 | total[b] | norm.[c] | rel.[d] |
| aug-cc-pV(Q+d)Z | 9129 | 4068 | 3218 | 3218 | 2528 | 22161 | 5.7 | 1.5 |
| jul-cc-pV(Q+d)Z | 7865 | 3460 | 3058 | 3058 | 2224 | 19665 | 5.1 | 1.3 |
| jun-cc-pV(Q+d)Z | 7262 | 3199 | 2770 | 2770 | 2044 | 18045 | 4.6 | 1.2 |
| may-cc-pV(Q+d)Z | 6793 | 2996 | 2546 | 2546 | 1904 | 16785 | 4.3 | 1.1 |
| apr-cc-pV(Q+d)Z | 6458 | 2851 | 2386 | 2386 | 1804 | 15885 | 4.1 | 1.0 |
| cc-pV(Q+d)Z | 6190 | 2735 | 2258 | 2258 | 1724 | 15165 | 3.9 | 1.0 |
| aug-cc-pV(T+d)Z | 5034 | 2208 | 1900 | 1900 | 1411 | 12453 | 3.2 | 1.5 |
| jul-cc-pV(T+d)Z | 4323 | 1866 | 1810 | 1810 | 1240 | 11049 | 2.8 | 1.4 |
| jun-cc-pV(T+d)Z | 3854 | 1663 | 1586 | 1586 | 1100 | 9789 | 2.5 | 1.2 |
| may-cc-pV(T+d)Z | 3519 | 1518 | 1426 | 1426 | 1000 | 8889 | 2.3 | 1.1 |
| cc-pV(T+d)Z | 3251 | 1402 | 1298 | 1298 | 920 | 8169 | 2.1 | 1.0 |
| aug-cc-pV(D+d)Z | 2387 | 1009 | 1024 | 1024 | 685 | 6129 | 1.6 | 1.6 |
| jul-cc-pV(D+d)Z | 2071 | 857 | 984 | 984 | 609 | 5505 | 1.4 | 1.4 |
| jun-cc-pV(D+d)Z | 1736 | 712 | 824 | 824 | 509 | 4605 | 1.2 | 1.2 |
| cc-pV(D+d)Z | 1468 | 596 | 696 | 696 | 429 | 3885 | 1.0 | 1.0 |

[a] Number of contracted basis functions used for all of the calculations in the database. [b] Sum of $N_{bf}$ over all five databases. [c] Sum normalized to cc-pV(D+d)Z. [d] Sum relative to cc-pV(*n*+d)Z for the same *n*.

***Table 3.*** Timing Summary for MP2 and MP2-F12 for All Species in the Databases

| | MP2 | | MP2-F12 | | |
|---|---|---|---|---|---|
| basis set[a] | norm.[b] | rel.[c] | norm. (MP2)[b] | norm.[d] | rel.[e] |
| aug-QZ | 388.9 | 6.3 | 942.6 | 46.9 | 3.4 |
| jul-QZ | 154.4 | 2.5 | 492.4 | 24.5 | 1.8 |
| jun-QZ | 112.9 | 1.8 | 402.7 | 20.0 | 1.4 |
| may-QZ | 88.3 | 1.4 | 337.9 | 16.8 | 1.2 |
| apr-QZ | 72.3 | 1.2 | 303.0 | 15.1 | 1.1 |
| QZ | 62.1 | 1.0 | 278.4 | 13.9 | 1.0 |
| aug-TZ | 30.1 | 5.0 | 144.2 | 7.2 | 2.8 |
| jul-TZ | 15.6 | 2.6 | 80.3 | 4.0 | 1.5 |
| jun-TZ | 10.4 | 1.7 | 65.7 | 3.3 | 1.3 |
| may-TZ | 7.8 | 1.3 | 57.8 | 2.9 | 1.1 |
| TZ | 6.0 | 1.0 | 51.9 | 2.6 | 1.0 |
| aug-DZ | 2.4 | 2.4 | 38.1 | 1.9 | 1.9 |
| jul-DZ | 1.8 | 1.8 | 24.8 | 1.2 | 1.2 |
| jun-DZ | 1.3 | 1.3 | 22.1 | 1.1 | 1.1 |
| DZ | 1.0 | 1.0 | 20.1 | 1.0 | 1.0 |

[a] Labels month-*n*Z or *n*Z denote month-cc-pV(*n*+*d*)Z or cc-pV(*n*+*d*)Z. [b] Sum normalized to MP2 with the same program on the same machine with the smallest basis set used (cc-pV(D+d)Z). [c] Sum relative to MP2 with the same program on the same machine with the smallest basis set with the same *n*. [d] Sum relative to MP2-F12 with the same program on the same machine with the smallest basis set used. [e] Sum relative to MP2-F12 with the same program on the same machine with the smallest basis set with the same *n*.

the HB4 database. In comparison with the results given in Table 4, the counterpoise-corrected results do not show different trends or lead to different conclusions on augmentation, and the errors are increased upon making the "correction". Therefore, to make the rest of the convergence tests as straightforward as possible, no counterpoise corrections are applied in Tables 4 or 6−8.

## 5. Discussion

In the Discussions and Conclusions (as in Table 3), we abbreviate cc-pV(D+d)Z, cc-pV(T+d)Z, and cc-pV(Q+d)Z as DZ, TZ, and QZ, respectively. Similarly, *month*-cc-pV(*n*+d)Z is abbreviated in the text as *month*-*n*Z.

Tables 4−6 show the mean deviations of MP2 energies computed with augmented, partially augmented, and not augmented basis sets from the near-CBS reference values. In all cases, for triple- and quadruple-ζ basis sets, the difference in MUD for jul- and jun- basis sets compared to aug- is usually small compared to the absolute value of the deviations, and in most cases they are negligibly small. The same holds true for jul-DZ.

The tables show many outstanding successes of partial augmentation, especially for barrier heights and hydrogen bond energies. For example, for barrier heights, aug-QZ, apr-QZ, and QZ differ from the near-CBS reference value by 0.2, 0.3, and 1.2 kcal/mol, respectively, but Table 2 shows that aug-QZ has about 50% more basis functions than apr-QZ. A similar observation can be made for may-TZ. For hydrogen bond energies, apr-QZ, may-TZ, and jul-DZ are all more accurate than aug-QZ. *The key point is not that they are more accurate, which (obviously) arises mainly from cancellation of errors, but that they are not significantly less accurate.*

A comparison of Tables 4 and 6 to Tables 7 and 8 shows that MP2 results converge more slowly with respect to the saturation of the diffuse space of the basis set than do MP2-F12 calculations, as expected from the discussion in section 1. Tables 7 and 8 show that in MP2-F12 calculations, only triple- and quadruple-ζ atomization energy calculations seem to be insensitive to the addition of diffuse basis functions. This is probably due to the fact that the underlying cc-pV(*n*+d)Z basis set is sufficiently diffuse for $n = \{T \text{ or } Q\}$ and not diffuse enough for $n = D$ for the calculation of bond energies. However, the inclusion of diffuse functions seems to be crucial for all other properties, including (perhaps surprisingly) ionization potentials.

It is interesting that for the tested properties the aug basis sets never offer any significant improvement at either the MP2 or MP2-F12 level over jul- basis sets. On the other hand, the unaugmented basis sets usually have much higher MUDs than partially augmented basis sets. Therefore, partially augmented basis sets provide intermediate options that are more balanced than either fully augmented or unaugmented basis sets with respect to the computational cost and the quality of the results.

***Table 4.*** Mean and Maximum Deviations of MP2 Barrier Height (kcal/mol), Hydrogen Bonding Energy (kcal/mol), and Electron Affinity (kcal/mol) from Reference Value

| | barrier height | | | hydrogen bonding | | | electron affinity | | |
|---|---|---|---|---|---|---|---|---|---|
| | MSD | MUD | MaxUD | MSD | MUD | MaxUD | MSD | MUD | MaxUD |
| aug-cc-pV(Q+d)Z | −0.03 | 0.15 | 0.43 | 0.09 | 0.09 | 0.13 | 0.64 | 1.36 | 2.64 |
| jul-cc-pV(Q+d)Z | 0.06 | 0.19 | 0.57 | 0.01 | 0.01 | 0.02 | 0.68 | 1.41 | 2.64 |
| jun-cc-pV(Q+d)Z | 0.13 | 0.20 | 0.61 | −0.01 | 0.02 | 0.03 | 1.14 | 1.70 | 3.18 |
| may-cc-pV(Q+d)Z | 0.20 | 0.25 | 0.78 | −0.03 | 0.03 | 0.04 | 1.50 | 1.93 | 3.49 |
| apr-cc-pV(Q+d)Z | 0.20 | 0.29 | 0.89 | 0.01 | 0.02 | 0.04 | 1.97 | 2.28 | 3.93 |
| cc-pV(Q+d)Z | −0.40 | 1.18 | 8.27 | 0.38 | 0.38 | 0.49 | 7.18 | 7.18 | 3.18 |
| aug-cc-pV(T+d)Z | 0.01 | 0.49 | 1.05 | 0.17 | 0.17 | 0.21 | 1.90 | 2.28 | 4.29 |
| jul-cc-pV(T+d)Z | 0.19 | 0.58 | 1.39 | 0.04 | 0.04 | 0.08 | 2.01 | 2.38 | 4.29 |
| jun-cc-pV(T+d)Z | 0.49 | 0.66 | 2.14 | −0.03 | 0.04 | 0.07 | 3.04 | 3.10 | 5.60 |
| may-cc-pV(T+d)Z | 0.65 | 0.79 | 3.05 | 0.05 | 0.09 | 0.10 | 4.47 | 4.47 | 6.96 |
| cc-pV(T+d)Z | −0.18 | 2.15 | 11.10 | 0.88 | 0.88 | 1.05 | 14.83 | 14.83 | 32.17 |
| aug-cc-pV(D+d)Z | −0.28 | 1.21 | 3.25 | 0.34 | 0.34 | 0.48 | 4.09 | 4.64 | 8.95 |
| jul-cc-pV(D+d)Z | 0.20 | 1.44 | 4.41 | 0.10 | 0.10 | 0.19 | 4.36 | 4.91 | 8.95 |
| jun-cc-pV(D+d)Z | 1.27 | 2.16 | 5.50 | 0.31 | 0.31 | 0.45 | 8.94 | 8.94 | 13.36 |
| cc-pV(D+d)Z | 0.08 | 3.92 | 15.81 | 1.93 | 1.93 | 2.44 | 30.26 | 30.26 | 61.09 |

***Table 5.*** Mean and Maximum Deviations of MP2 Counterpoise-Corrected Hydrogen Bonding Energy (kcal/mol) from the Reference Value

| | hydrogen bonding | | |
|---|---|---|---|
| | MSD | MUD | MaxUD |
| aug-cc-pV(T+d)Z | −0.24 | 0.24 | 0.28 |
| jul-cc-pV(T+d)Z | −0.29 | 0.29 | 0.36 |
| jun-cc-pV(T+d)Z | −0.38 | 0.38 | 0.45 |
| may-cc-pV(T+d)Z | −0.45 | 0.45 | 0.56 |
| cc-pV(T+d)Z | −0.51 | 0.51 | 0.81 |

***Table 6.*** Mean and Maximum Deviations of MP2 Ionization Potentials (kcal/mol) and Atomization Energies per Bond (kcal/mol per Bond) from Reference Value

| | ionization potential | | | atomization energy | | |
|---|---|---|---|---|---|---|
| | MSD | MUD | MaxUD | MSD | MUD | MaxUD |
| aug-cc-pV(Q+d)Z | −2.00 | 2.26 | 7.52 | −1.03 | 1.03 | 1.82 |
| jul-cc-pV(Q+d)Z | −2.02 | 2.27 | 7.52 | −1.09 | 1.09 | 1.97 |
| jun-cc-pV(Q+d)Z | −2.27 | 2.47 | 7.86 | −1.21 | 1.21 | 2.12 |
| may-cc-pV(Q+d)Z | −2.38 | 2.55 | 8.09 | −1.26 | 1.26 | 2.18 |
| apr-cc-pV(Q+d)Z | −2.43 | 2.60 | 8.19 | −1.28 | 1.28 | 2.19 |
| cc-pV(Q+d)Z | −2.55 | 2.70 | 7.86 | −1.28 | 1.28 | 2.17 |
| aug-cc-pV(T+d)Z | −3.26 | 3.28 | 7.80 | −2.44 | 2.44 | 4.28 |
| jul-cc-pV(T+d)Z | −3.28 | 3.30 | 7.80 | −2.60 | 2.60 | 4.68 |
| jun-cc-pV(T+d)Z | −3.89 | 3.89 | 8.67 | −2.86 | 2.86 | 5.08 |
| may-cc-pV(T+d)Z | −4.20 | 4.20 | 9.17 | −2.98 | 2.98 | 5.19 |
| cc-pV(T+d)Z | −4.55 | 4.55 | 10.40 | −3.01 | 3.01 | 5.09 |
| aug-cc-pV(D+d)Z | −6.25 | 6.25 | 10.99 | −7.38 | 7.38 | 13.87 |
| jul-cc-pV(D+d)Z | −6.31 | 6.31 | 10.99 | −7.69 | 7.69 | 14.68 |
| jun-cc-pV(D+d)Z | −8.65 | 8.65 | 13.81 | −8.51 | 8.51 | 15.87 |
| cc-pV(D+d)Z | −10.23 | 10.23 | 19.13 | −8.62 | 8.62 | 15.57 |

At any given $\zeta$ level ($n$) and augmentation level (*month*-), if one can afford additional cost and wishes to improve the quality of the calculations by increasing the size of the basis set, one would only be interested in doing so if the larger basis set offered significantly lower mean deviations. Mean deviations in electron affinity listed in Table 7 show that only partially augmented basis sets offer such an advantage. To see this for MP2-F12 calculations, recall that in the tables each basis set is larger than all those below it (as shown quantitatively in Table 2). Then, consider starting with DZ and moving up. The error in Table 7 decreases significantly as we increase the basis set only through jun- and jul-, but not for aug-. The higher, triple-$\zeta$ unaugmented basis set also does not offer a decrease in error relative to that of jul-DZ. Therefore, aug-DZ and -TZ should be skipped. The next basis set to offer an improvement in the quality of the results is jun-TZ. The jul-TZ basis set offers further improvement. Then, aug-TZ and QZ can be skipped until improvement is again found with apr-QZ. Similar conclusions are drawn from other tables, for both MP2 and MP2-F12, except for atomization energies at the triple- and quadruple-$\zeta$ levels (Tables 6 and 8), which do not seem to require diffuse functions.

The trend is even more striking for calculations without F12. Consider, for example, the barrier height calculations in Table 4. One achieves higher accuracy with apr-QZ than with aug-TZ, jul-TZ, jun-TZ, or may-TZ, but simply going to QZ without any diffuse functions is less accurate than any of these triple-$\zeta$ levels. Table 3 shows that the cost savings in using apr-QZ rather than aug-QZ to include the diffuse space is more than a factor of 5. This is one of the main lessons of the present study and is more important than the small deviations from one MUD to another in the 0.5−0.8 kcal/mol range or the small deviations from one MUD to another in the 0.15−0.3 kcal/mol range. Similarly, now considering barrier heights in Table 4, TZ is inaccurate, but may-TZ is more accurate than aug-DZ and is less expensive than aug-TZ by about a factor of 4. Thus, in calculations where one cannot afford to go all the way to aug-TZ, the errors may be decreased considerably compared to TZ by using intermediate augmentation. As a third example, consider hydrogen bonding in Table 4. The main point is not the small differences from one MUD to another in the range 0.01−0.17 kcal/mol but rather the fact that even apr-QZ accounts well for the effect of diffuse functions at the quadruple-$\zeta$ level. Even may-TZ accounts well for them at the TZ level, and even jun-DZ accounts well for them at the DZ level. Usually, when diffuse functions are important, we find that ($n$+1)Z is not more accurate than jul-$n$Z or jun-$n$Z. In other words, the fully augmented and unaugmented basis sets hardly ever seem to be good choices as compared to the new intermediate basis set levels, and in almost all cases,

***Table 7.*** Mean and Maximum Deviations of MP2-F12 Barrier Height (kcal/mol), Hydrogen Bonding Energy (kcal/mol), and Electron Affinity from Reference Value

| | barrier height | | | | | | | hydrogen bonding | | | electron affinity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | | | MUD | | | | | | | | | |
| | DBH24 | (DBH06[a] | DBH18[b]) | DBH24 | (DBH06[a] | DBH18[b]) | MaxUD | MSE | MUD | MaxUD | MSD | MUD | MaxUD |
| aug-cc-pV(Q+d)Z | 0.00[c] | (0.00[c] | 0.00[c]) | 0.00[c] | (0.00[c] | 0.00[c]) | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] |
| jul-cc-pV(Q+d)Z | 0.02 | (0.00 | 0.02) | 0.03 | (0.02 | 0.03) | 0.08 | 0.02 | 0.02 | 0.03 | 0.03 | 0.04 | 0.13 |
| jun-cc-pV(Q+d)Z | 0.02 | (0.00 | 0.03) | 0.03 | (0.01 | 0.03) | 0.08 | 0.02 | 0.02 | 0.02 | 0.10 | 0.10 | 0.14 |
| may-cc-pV(Q+d)Z | 0.03 | (0.01 | 0.04) | 0.04 | (0.04 | 0.04) | 0.10 | 0.02 | 0.02 | 0.03 | 0.27 | 0.27 | 0.42 |
| apr-cc-pV(Q+d)Z | 0.00 | (−0.12 | 0.05) | 0.08 | (0.12 | 0.06) | 0.40 | 0.08 | 0.08 | 0.10 | 0.71 | 0.71 | 0.91 |
| cc-pV(Q+d)Z | −0.51 | (−2.07 | 0.00) | 0.87 | (3.04 | 0.15) | 7.09 | 0.46 | 0.46 | 0.57 | 4.88 | 4.88 | 13.63 |
| aug-cc-pV(T+d)Z | 0.06 | (−0.06 | 0.10) | 0.11 | (0.07 | 0.12) | 0.77 | 0.04 | 0.04 | 0.04 | 0.23 | 0.23 | 0.36 |
| jul-cc-pV(T+d)Z | 0.09 | (−0.04 | 0.14) | 0.16 | (0.09 | 0.18) | 0.69 | 0.14 | 0.14 | 0.15 | 0.33 | 0.33 | 0.56 |
| jun-cc-pV(T+d)Z | 0.13 | (0.03 | 0.16) | 0.18 | (0.14 | 0.20) | 0.75 | 0.14 | 0.14 | 0.15 | 0.62 | 0.62 | 1.07 |
| may-cc-pV(T+d)Z | 0.16 | (0.00 | 0.21) | 0.25 | (0.22 | 0.26) | 1.12 | 0.24 | 0.24 | 0.28 | 1.86 | 1.86 | 2.47 |
| cc-pV(T+d)Z | −0.56 | (−2.53 | 0.10) | 1.44 | (4.37 | 0.46) | 9.24 | 1.09 | 1.09 | 1.27 | 9.79 | 9.79 | 22.66 |
| aug-cc-pV(D+d)Z | 0.03 | (0.13 | 0.00) | 0.26 | (0.32 | 0.25) | 0.69 | 0.08 | 0.08 | 0.10 | 1.02 | 1.02 | 1.83 |
| jul-cc-pV(D+d)Z | 0.15 | (0.21 | 0.13) | 0.39 | (0.40 | 0.39) | 1.32 | 0.38 | 0.38 | 0.46 | 1.18 | 1.18 | 1.78 |
| jun-cc-pV(D+d)Z | 0.54 | (1.31 | 0.28) | 0.84 | (1.37 | 0.66) | 2.19 | 0.56 | 0.56 | 0.70 | 4.27 | 4.27 | 5.71 |
| cc-pV(D+d)Z | −0.54 | (−2.14 | 0.00) | 2.66 | (6.76 | 1.29) | 12.38 | 2.31 | 2.31 | 2.86 | 19.64 | 19.64 | 39.98 |

[a] Reactions containing anions ($Cl^-\cdots CH_3Cl$, $F^-\cdots CH_3Cl$, $OH^-\cdots CH_3F$). [b] The rest of the reactions in DBH24/08 (not listed in *a*). [c] Zero by definition.

***Table 8.*** Mean and Maximum Deviations of MP2-F12 Ionization Potentials (kcal/mol) and Atomization Energies per Bond (kcal/mol per Bond) from Reference Value

| | ionization potential | | | atomization energy | | |
|---|---|---|---|---|---|---|
| | MSD | MUD | MaxUD | MSD | MUD | MaxUD |
| aug-cc-pV(Q+d)Z | 0.00[a] | 0.00[a] | 0.00[a] | 0.00[a] | 0.00[a] | 0.00[a] |
| jul-cc-pV(Q+d)Z | −0.01 | 0.02 | 0.04 | 0.01 | 0.01 | 0.02 |
| jun-cc-pV(Q+d)Z | −0.04 | 0.04 | 0.06 | 0.00 | 0.01 | 0.01 |
| may-cc-pV(Q+d)Z | −0.09 | 0.09 | 0.13 | −0.01 | 0.01 | 0.02 |
| apr-cc-pV(Q+d)Z | −0.15 | 0.15 | 0.23 | −0.01 | 0.01 | 0.03 |
| cc-pV(Q+d)Z | −0.31 | 0.31 | 0.90 | 0.01 | 0.02 | 0.05 |
| aug-cc-pV(T+d)Z | −0.13 | 0.14 | 0.26 | −0.03 | 0.04 | 0.06 |
| jul-cc-pV(T+d)Z | −0.17 | 0.17 | 0.25 | −0.01 | 0.03 | 0.06 |
| jun-cc-pV(T+d)Z | −0.35 | 0.35 | 0.58 | −0.03 | 0.04 | 0.08 |
| may-cc-pV(T+d)Z | −0.68 | 0.68 | 1.02 | −0.04 | 0.04 | 0.11 |
| cc-pV(T+d)Z | −1.28 | 1.28 | 2.77 | 0.06 | 0.07 | 0.18 |
| aug-cc-pV(D+d)Z | −0.74 | 0.80 | 1.50 | −0.19 | 0.22 | 0.46 |
| jul-cc-pV(D+d)Z | −0.99 | 1.01 | 1.62 | −0.10 | 0.19 | 0.33 |
| jun-cc-pV(D+d)Z | −2.86 | 2.86 | 5.36 | −0.06 | 0.23 | 0.35 |
| cc-pV(D+d)Z | −4.91 | 4.91 | 9.08 | 0.24 | 0.34 | 0.87 |

[a] Zero by definition.

truncation to a basis set intermediate between unaugmented or fully augmented on only heavy atoms is the best choice.

An excessive number of diffuse functions not only increases the cost but also often leads to difficult SCF convergence, and as shown in Figure 1, it can increase the basis set superposition error.

The fact that the aug-*n*Z basis sets approach the CBS limit in a systematic way is often correctly considered to be one of their chief advantages. However, if one accepts that diffuse functions are not needed on hydrogenic atoms, a basis set sequence such as jun-XZ ($x$ = D, T, Q, etc.) also approaches the CBS limit systematically (i. e., is "convergent"), and it does so in a more efficient manner. Thus, partially augmented basis sets would appear to be very useful for focal point[23,64] analysis.

We have not examined reduction in the number of polarization functions, but previous work[65] shows that savings are possible in that area as well.

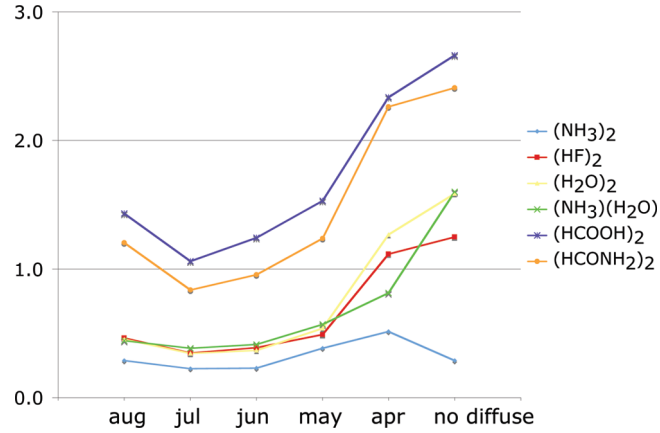



***Figure 1.*** Counterpoise corrections [kcal/mol] for hydrogen-bonded dimers for triple-ζ basis sets.

Another possible use of the partially augmented basis sets is in dual-basis calculations where Hartree−Fock calculations are performed in a small basis set and the post-Hartree−Fock calculation is performed in a large set.[66,67]

The second objective of this article is to compare MP2-F12 to MP2. Both kinds of calculations converge to the same MP2 CBS limit, but at different rates. Comparing Table 4 to Table 7 shows that, for barrier heights, MP2-F12/jul-DZ is more accurate than MP2/aug-TZ or MP2/aug-QZ. Furthermore, MP2-F12/may-TZ is as accurate as MP2/jul-QZ. For ionization potentials (Tables 6 and 8), MP2-F12/jul-DZ is more accurate than MP2/jun-QZ. For atomization energies (Tables 6 and 8), we find that MP2-F12/DZ is more accurate than MP2/aug-QZ. However, MP2-F12 is not more accurate than MP2 for hydrogen bonding.

## 6. Conclusions

Of all of the energetic molecular properties considered here, unaugmented basis sets are adequate only for atomization energies at the triple- and quadruple-ζ levels.

Both MP2 and MP2-F12 theories are sensitive to the saturation of the diffuse space; however, in most cases

presented here, the full augmentation is unnecessary. Instead of using fully augmented (aug) or unaugmented basis sets, we recommend using the new partially augmented basis sets.

For MP2 calculations of properties requiring diffuse basis functions, we recommend jun-QZ, jul-TZ, and jul-DZ for especially reliable results, but the tables show that, except for electron affinities (and therefore probably for most properties involving anions), one can usually cut back to may-QZ or jun-TZ.

In MP2-F12 calculations of properties sensitive to the number of diffuse functions, we recommend using may-QZ, jun-TZ, and jul-DZ, which offer considerable savings compared to aug- basis sets and significant improvement over unaugmented cc-pV($n$+d)Z basis sets.

**Acknowledgment.** We thank Dr. Steven Mielke for helpful discussions and Dr. Boris Averkiev for help with the MP2-F12 calculations. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, under grant no. DE-FG02-86ER13579. The MP2-F12 part of this research was performed using the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory, operated for the Department of Energy by Battelle. Resources for MP2 calculations were provided by Minnesota Supercomputing Institute.

**Supporting Information Available:** Root mean square errors for all the properties in this article. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657.

(2) Davidson, E. R.; Feller, D. *Chem. Rev.* **1986**, *86*, 681.

(3) Clark, T.; Chandresekhar, J.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

(4) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(5) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975.

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

(7) Koput, J.; Peterson, K. A. *J. Phys. Chem. A* **2002**, *106*, 9595.

(8) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.

(9) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **1999**, *110*, 7667.

(10) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.

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

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

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

(14) Del Bene, J. E.; Shavitt, I. *THEOCHEM* **1994**, *307*, 27.

(15) Lynch, B. J. Truhlar, D. G. In *Electron Correlation Methodology*; Wilson, A. K., Peterson, K. A. Eds.; ACS Symposium Series 958; American Chemical Society: Washington, DC, 2007; p 153.

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

(17) Papajak, E.; Truhlar, G. D. *J. Chem. Theory Comput.* **2010**, *6*, 597–601.

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

(19) Petersson, G. A.; Braunstein, M. *J. Chem. Phys.* **1985**, *83*, 5129.

(20) East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638.

(21) Hobza, P.; Sponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802.

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

(23) Schuurmann, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *120*, 11586.

(24) Soteras, I.; Orozco, M.; Luque, F. J. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 281.

(25) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 114104.

(26) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478.

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

(28) Fast, P. L.; Corchado, J. C.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 5129.

(29) Peterson, K. A.; Dunning, T. H. *J. Phys. Chem.* **1995**, *99*, 3898.

(30) Schwenke, D. *J. Chem. Phys.* **2005**, *122*, 014107.

(31) Kutzelnigg, W. *Theor. Chim. Acta* **1985**, *68*, 445.

(32) Kutzelnigg, W.; Klopper, W. *J. Chem. Phys.* **1991**, *94*, 1985.

(33) Klopper, W.; Kutzelnigg, W. *Chem. Phys. Lett.* **1987**, *134*, 17.

(34) Klopper, W.; Samson, C. C. M. *J. Chem. Phys.* **2002**, *116*, 6397.

(35) Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 4607.

(36) Ten-no, S.; Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 5358.

(37) Ten-no, S. *J. Chem. Phys.* **2004**, *121*, 117.

(38) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190.

(39) Valeev, E. F.; Jansen, C. L. *Chem. Phys. Lett.* **2004**, *121*, 1214.

(40) Ten-no, S. *Chem. Phys. Lett.* **2004**, *121*, 117.

(41) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.

(42) Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610.

(43) Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.

(44) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580.

(45) Bischoff, F. A.; Wolfsegger, S.; Tew, D. P.; Klopper, W. *Mol. Phys.* **2009**, *107*, 963.

(46) Werner, H.-J.; Knizia, G.; Adler, T. B.; Marchetti, O. *Z. Phys. Chem.* **2010**, *224*, 493.

(47) Lane, J. R.; Kjaergaard, H. G. *J. Chem. Phys.* **2009**, *131*, 034307.

(48) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.

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

(50) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(51) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Koppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pfluger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2009.1; University College Cardiff Consultants Limited: Cardiff, United Kingdom.

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

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

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

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

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

(57) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *107*, 8996.

(58) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285.

(59) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59.

(60) Weigend, F.; Kohn, A.; Hattig, C. *J. Chem. Phys.* **2002**, *116*, 3175.

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

(62) Alvarez-Idaboy, J. R.; Galano, A. *Theor. Chem. Acc.* **2010**, *126*, 75.

(63) Skwara, B.; Bartkowiak, W.; Da Silva, D. L. *Theor. Chem. Acc.* **2009**, *122*, 127.

(64) King, R. A.; Allen, W. D.; Ma, B.; Schaefer, H. F., III. *Faraday Discuss.* **1998**, *110*, 23.

(65) Mintz, B.; Lennox, K. P.; Wilson, A. K. *J. Chem. Phys.* **2004**, *121*, 5629.

(66) Jurgens-Lutovsky, R.; Almlof, J. *J. Chem. Phys. Lett.* **1991**, *178*, 431.

(67) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2003**, *118*, 9497.

CT1005533

JCTC Journal of Chemical Theory and Computation

# Correcting Systematic Errors in DFT Spin-Splitting Energetics for Transition Metal Complexes

Thomas F. Hughes and Richard A. Friesner*

*Columbia University, Department of Chemistry, New York, New York 10027, United States*

Received June 28, 2010

**Abstract:** Spin-splittings of 57 octahedral first-row transition metal complexes calculated with B3LYP are compared with a database of experimental spectra collected from the literature. A variety of transition metal centers in various oxidation states and multiplicities along with a number of different coordinating ligands are considered. Environmental effects have been included to enable reasonable quantitative comparison with experiment. The manifold of states is studied using initial guesses constructed from ligand field theory. A localized orbital correction (LOC) model, referred to as DBLOC-DFT (d-block localized orbital corrected density functional theory), systematically corrects B3LYP calculations using five parameters. The final results are a considerable improvement over conventional DFT, bringing the mean unsigned error (MUE) from 10.14 kcal/mol with a standard deviation of 4.56 to 1.98 kcal/mol with a standard deviation of 1.62. Depending on the relative multiplicities of the ground and excited states, it is shown that B3LYP*, which has 15% exact nonlocal exchange, can lead to larger errors with respect to experiment than B3LYP. Application to 7 complexes from Swart et al. [*J. Phys. Chem. A* **2004**, *108*, 5479.] and 14 small-gap spin-crossover complexes, from the literature, shows the DBLOC model provides good agreement with a variety of experimental data.

## 1. Introduction

Transition metals play very important roles in biochemistry and materials science. The catalytic properties of metalloenzymes and other metal containing catalysts are a result of special electronic structures of reactive intermediates. As examples, consider the catalytic sites in methane monooxygenase[1] and the blue ruthenium dimer[2] for which the catalytically competent species have unusually large metal oxidation states along with important magnetic coupling interactions. The high density of low lying electronic states resulting from the metal atom can give rise to multiple reactive channels that can utilize spin-crossover[3−6] mechanisms. Here, otherwise high energy barriers become low energy through a facile change in spin multiplicity. Qualitatively, spin-crossover becomes possible when there is a balance between the pairing energy of two electrons and the ligand field splitting energy, $\Delta_o$, needed to traverse the $t_{2g} \rightarrow e_g$ gap.[3] Splitting energies depend on metal−ligand interactions which range from weak interactions like simple electrostatically perturbed lone pairs to stronger dative bonds and finally more covalent bonds. In principle, these can be systematically tuned so as to control relative spin state energetics.

The energy difference between the $t_{2g}$ and $e_g$ sets of orbitals, $\Delta_o$, has been found empirically to depend on the coordinating ligands according to the spectrochemical series[7−11]

$$I^{-1} < Br^{-1} < S^{-2} < NCS^{-1} < Cl^{-1} < NO_3^{-1} < N_3^{-1} < F^{-1} < OH^{-1} < C_2O_4^{-2} < H_2O < SCN^{-1} < CH_3CN < \text{pyridine} < NH_3 < \text{en} < \text{bipy} < \text{phen} < NO_2^{-1} < SO_3^{-2} < PPh_3 < CN^{-1} < CO$$

as well as the oxidation state of the metal center and the identity of the metal itself. Moving right in the series increases $\Delta_o$, while decreasing the oxidation state of the metal decreases $\Delta_o$ so the balance of electronic effects becomes important. A large value of $\Delta_o$ signifies strong metal−ligand

* To whom correspondence should be addressed. E-mail: rich@chem.columbia.edu.

10.1021/ct100359x 
Published on Web 12/07/2010

coupling, thereby raising the antibonding $e_g$ orbital energies, oftentimes resulting in a low-spin ground state where the complex would rather pair electrons than spend additional energy in traversing the gap. Conversely, small values of $\Delta_o$ typically lead to high-spin complexes.

Requiring an electronic structure method to be capable of correctly predicting transition metal chemistry along with exhibiting calculation speed capable of handling large systems is a demanding request. It is well-known that the computation of a property like the spin-splitting energy is very sensitive to the quality of the underlying wave function. Two separate questions to answer are: what is the ground state multiplicity, and what is the energy for traversing the spin gap? Inclusion of factors like multireference character, dynamic correlation, spin−orbit coupling, relativity, vibronic coupling, as well as antiferromagnetic or ferromagnetic coupling for the case of transition metal clusters would be required for very high-level calculations. High-level methods, for example multireference many-body methods like coupled-cluster theory for transition metal systems, in principle offer a solution to the electronic problem, however, their computational cost, outside of massively parallel applications for which system size is still relatively small, renders it less useful in materials science and biology than density functional theory (DFT).[12−14]

Hybrid DFT methods such as B3LYP, the principal functional investigated in this work, provide an alternative to accurate, but computationally expensive, wave function-based approaches. Because even high-level approaches, such as coupled-cluster with perturbative triples (CCSD(T)), can be problematic for calculations involving transition metals,[15] DFT approaches, which for many cases deliver reasonable results, at present must be considered the method of choice for these systems. However, over the past decade, significant evidence has accumulated that even the best current DFT functionals can display large errors for specific types of transition metal energetics. As an example, different functionals can yield very different answers for spin state splittings by as much as 10−30 kcal/mol, and such errors can be manifested in comparisons with experimental data, for example in predicting the wrong ground state or in failure to properly yield small splittings in finely tuned spin-crossover complexes.

Despite the above observations, a detailed and comprehensive understanding of the performance of B3LYP, or any other functional, for spin state energetics is not currently available, in great part, because of the paucity and problematic nature of the available experimental data. Hence, as an initial step, we have constructed a chemically diverse database of solution and crystalline experimental spectra of 57 octahedral first-row transition metal containing complexes with different properties. Our database consists of 2 V,[16,17] 10 Cr,[7,17−23] 6 Mn,[23−26] 23 Ni,[22,27−34] 7 Fe,[35−39] and 9 Co[40−45] containing complexes with +2, +3, and +4 oxidation states, total charges ranging from −4 to +4, multiplicities ranging from 1 to 6, and total number of valence electrons ranging from 2 to 8. Coordinating atoms include C, N, O, S, F, and Cl with differing partial atomic charges, and the smallest and largest ligands contain 2 and 96 atoms, respectively. Ligands vary from mono- to hexa-dentate and in all cases are lone pair coordinated, i.e. metallocenes and other complexes with high hapticity are not considered here.

The variation in this database should be contrasted with alternative data sources such as electron paramagnetic resonance (EPR) or optical absorption measurements of spin-crossover complexes, where as has recently been pointed out by Deeth et al.,[46] the vast majority of complexes have aromatic nitrogens coordinated to Fe(II), and the ground states are uniformly low-spin singlets, with the excited state quintet requiring promotion of two electrons to a higher lying d-orbital, for example $t_{2g} \rightarrow e_g$. The spin-crossover data is useful but the drastic limitations noted above, the significance of which will become more apparent as we proceed with our analysis, has not been generally recognized.

To isolate the errors in the electronic problem from any potential errors because of the strong vibronic coupling effects and nonradiative transitions present at the spin-crossing seam, our database is based on equilibrium radiative spin-forbidden transitions under the Born−Oppenheimer approximation. Adiabatic potential energy surfaces are adequate because only regions that are far from the crossing are studied. These transitions are singly forbidden (the change in the number of unpaired electrons in going to the excited state is always two) d−d transitions that violate the spin selection rule and so they are of low intensity. Typically they show up in the optical absorption spectrum of the complex as shoulders to more intense spin allowed metal-to-ligand charge transfer (MLCT) or ligand-to-metal charge transfer (LMCT) transitions. Forbidden transitions become more intense by borrowing intensity from allowed transitions through spin−orbit coupling.

With a large, diverse database in hand, we observe large, but systematic, errors in the B3LYP computation of transition metal spin state splittings. However, these errors do not follow a pattern as simplistic as has been suggested in most DFT studies of these splittings to date.[47,48] While it is true that for the great majority of the aromatic nitrogen dominated spin-crossover complexes noted above (on which previous analysis has been primarily based), B3LYP shows an overstabilization of the high-spin state resulting in it being the ground state. But as is shown in detail below, other types of complexes can display very different behavior; for a substantial number of cases, we find in fact that low-spin states are very substantially overstabilized. A consequence of considering this much broader range of experimental systems is that solutions to the DFT spin-splitting problem proposed previously, such as decreasing the amount of exact nonlocal exchange in the hybrid functional, no longer yield even qualitatively satisfactory results. Nevertheless, striking regularities in the data are manifested, and a satisfactory model containing a modest number of parameters, all of which are physically based, can be constructed.

Most new DFT functional development has involved optimization of parameters internal to the DFT energy expression, as well as addition of new terms such as those involving the kinetic energy.[49,50] This approach has proved fruitful with new functionals such as the meta-GGA functional series of Truhlar and co-workers displaying substan-

tially improved prediction capabilities for a variety of properties.[50] A second approach, specific to transition metals, involves the adjustment of the amount of exact nonlocal exchange in hybrid functionals, for example the B3LYP* functional reduces the amount of exact nonlocal exchange in B3LYP from 20% to 15%.[48,51] This functional has been calibrated primarily with spin-crossover data, which, as discussed above, is apparently problematic.

Recently, we have introduced a simple, yet highly effective alternative to these approaches, in which localized, valence-bond-type empirical corrections are applied to specific chemical groups, fitting the parameter values to experimental data.[15,52−55] These corrections represent an improved estimation of nondynamical correlation effects in the targeted bonds and lone pairs, which are assumed to be transferable for well-defined localized chemical functionalities. Near-chemical accuracy has been achieved for molecules containing second and third-row atoms for thermodynamics, barrier heights, and electron affinities and ionization potentials, for large and diverse databases.[15,52−55] An initial paper extending this approach to transition metals was published several years ago[15] and yielded substantial improvements in energetics for atoms and small transition metal species, primarily a metal atom bonded to one small ligand. However, in that work we did not consider realistic organometallic complexes, such as those that comprise the present database.

In the present paper, we show that a suitably simple empirical correction scheme, entirely consistent with our previous work, and containing only five parameters which are readily physically interpretable, yields a dramatic reduction in the mean unsigned error (MUE) of the computed spin state gaps compared to experiment. The accuracy, MUE of around 2 kcal/mol, is consistent with that obtained in our previous transition metal work, and the MUE is somewhat higher than that for organic systems (MUE of around 1 kcal/mol) principally because of the lack of sufficient experimental data and the noise inherent in that data.

## 2. Methods

For many complexes we used published X-ray crystal structures for the initial geometries. In cases where these were not available we sometimes found a crystal structure for the complex with a different metal center or built the complex from crystal structures of multiligand complexes where only one ligand was of interest. The Cambridge Structural Database[56] was used. Especially for nickel-containing complexes, for reasons discussed below, it was sometimes necessary to use the minimized ground state geometry as opposed to the crystal structure as the initial structure for the excited state geometry minimizations.

All calculations were done using Jaguar version 7.5[57] with the relativistic effective core potential LACV3P, a triple-$\zeta$ contraction of the LACVP[58] basis set, for metal centers and 6-311G for the rest of the atoms. This basis set was chosen because it was successfully used in developing a localized orbital correction (LOC) model for transition metal atoms[15] and has been shown in the literature to work well in many standard DFT applications to metal containing systems. Previous work has shown that while the LOC corrections for metal atoms[15] do show some basis set dependence the differences in average error of the optimized model between the medium (LACV3P) and very large (QZVP(-g)) basis sets are small considering the great variety in the transition metal database.

Adiabatic spin state energy gaps are calculated using fully unrestricted pseudospectral B3LYP, with a continuum dielectric employed to model environmental effects. The Poisson−Boltzmann solver in Jaguar[57] was used. We used dielectric constant values for whatever solvent was used in obtaining the spectra experimentally, for example water has a value of 80.4, or in the case of crystalline spectra a low dielectric constant of 2.0 is used. Zero-point energy and entropic corrections were not determined because their difference with respect to states of different spin multiplicities are typically small.

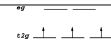
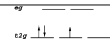
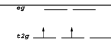
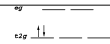
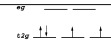
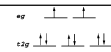
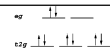
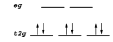
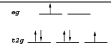
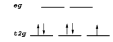
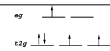
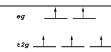
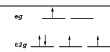
The method of Langlois and co-workers[59] was used to generate initial guess density matrices, which take into account d−d electron repulsion and metal−ligand interactions. These guesses help to alleviate some of the difficulties that quantum chemical methods have with convergence to the ground states of the various multiplicities in transition metal containing systems. To ensure a more thorough search of state space we construct a series of initial guesses given a user specified table of atomic formal charges, localized multiplicities, and in the case of ferromagnetic or antiferromagnetic coupling, the number of coupled alpha- and beta-spins. In the case of magnetic coupling, broken symmetry DFT was used.[60] The default value of the relative energy threshold (keyword `opt325`) for acceptable initial guess electron configurations was changed from 0.1 hartree to 1.0 hartree to explore each of the different states generated by permuting single and paired electrons among the d-orbitals. Note that there were many times when the initial guess energy ordering of states, as well as obviously their relative magnitudes, was different from the converged UDFT values because of qualitative differences in the density matrices. Usually for calculations using the ground state multiplicity the lowest energy states are obtained using low values of the istate keyword, however, there were some exceptions. For calculations using the excited state multiplicities, this was certainly not true. Therefore the full range of `istate` values was considered in each case; this range can be precomputed using simple combinatorics

$$\texttt{istate} \in \left[1, \frac{5!}{s!d!(5-s-d)!}\right] \qquad (1)$$

where $s$ and $d$ are the number of singly and doubly occupied orbitals respectively in the set of five d-orbitals for the complex of a given multiplicity. The `istate` keyword simply specifies an initial guess for the electronic configuration of the metal complex in terms of singly and doubly occupied d-orbitals.

Adiabatic spin-forbidden transition energies are calculated using the lowest energy state of each spin multiplicity. To compare the calculated adiabatic transition energies with experimental vertical transitions the vibrational relaxation energies must be estimated. Franck−Condon factors could be calculated,[61] and the vibrational relaxation energy deter-

***Table 1.*** Molecular Orbital Diagrams for Octahedral Metal Complexes with Spin-Forbidden Transitions

| complex | type | num. val. elec. | g.s. mult. | g.s. l.f.d. | e.s. mult. | e.s. l.f.d. |
|---|---|---|---|---|---|---|
| cr223tetcl2, cracac2im2py, crccsime36, crcn6, crcyclamncs2, cren3, crf6, crnh34cl2, crnh36, crox3, mnf6 | $t_{2g} \rightarrow t_{2g}$ | 3 | 4 |  | 2 |  |
| crf6, vf6, vurea6 | $t_{2g} \rightarrow t_{2g}$ | 2 | 3 |  | 1 |  |
| mn2p2pameth2, mncn6 | $t_{2g} \rightarrow t_{2g}$ | 4 | 3 |  | 1 |  |
| ni12dimeim6, ni1meim6, ni2meim6, niacac2im2py, nibipy3, nibpm2no3, nidmso6, nidpdpm2h2o2, nidpdpm2no3h2o, nidpdpmno32, nidpdpmno32ch3cn, niedta, nien2scn2, nien3, nif6, nigly3, nih2o6, ninh36, niphen3, nipyrazole6, nitach3mepyr, nitpm2, nitpmno32 | $e_g \rightarrow e_g$ | 8 | 3 |  | 1 |  |
| coamn3s3sarh, coamn5ssarh, coen3, coetn4s2amp, col2, con3s3, conh35soch32, conh36, fecn6 | $t_{2g} \rightarrow e_g$ | 6 | 1 |  | 3 |  |
| cof6 | $e_g \rightarrow t_{2g}$ | 7 | 4 |  | 2 |  |
| fecn6, feen3 | $t_{2g} \rightarrow e_g$ | 5 | 2 |  | 4 |  |
| fecat3, feeta3, feh2o6, fethiocarbamate3, fetrencam, mnden2, mnen3, mnh2o6 | $e_g \rightarrow t_{2g}$ | 5 | 6 |  | 4 |  |

mined as a sum of vibrational energies up to the state with the most intense transition; however, we choose to take the difference between peak and onset values of the spin-forbidden shoulder from the experimental spectra so as to avoid, among other potential complications, having to calculate intensity contributions from the environment. Direct calculation of vertical transition energies is avoided because they would, at least in certain cases where the equilibrium geometries of the two states are significantly different, introduce additional errors inherent in nonequilibrium UDFT calculations. Additionally, some SCF convergence issues were found for vertical calculations. The protocol we have adopted enables a direct comparison between theory and experiment for the energy gaps between different spin states.

## 3. Results and Discussion

**3.1. Database of $t_{2g} \rightarrow t_{2g}$, $e_g \rightarrow e_g$, $t_{2g} \rightarrow e_g$, and $e_g \rightarrow t_{2g}$ Spin-Forbidden Transitions for Octahedral Transition Metal Complexes.** Table 1 shows the octahedral metal complexes in the DBLOC (d-block localized orbital corrected) database, including $t_{2g} \rightarrow t_{2g}$, $e_g \rightarrow e_g$, $t_{2g} \rightarrow e_g$, and $e_g \rightarrow t_{2g}$ ground to spin-forbidden excited state transitions. Table 1 of the Supporting Information shows models of the $t_{2g} \rightarrow t_{2g}$ complexes along with the names to which they will be referred (see Supporting Information Table 4 for a list of descriptors). There are a total of 16 complexes, 10 Cr(III), 1 Cr(IV), 2 Mn(III), 1 Mn(IV), and 2 V(III), of this type. From Table 1 note that crf6 is listed twice, once as $Cr(III)F_6^{-3}$ and once as $Cr(IV)F_6^{-2}$. The molecular orbital diagrams are shown in Table 1. These were determined using a combination of experimental results, ligand field theory, Mulliken spin densities and populations, and natural electron configurations.[62] A more detailed example of a molecular orbital diagram from hybrid DFT calculations on octahedral transition metal complexes was recently given by Watts et al.[63] It can be seen that energy differences between orbitals belonging to a given manifold, $t_{2g}$ or $e_g$, are generally small. All complexes have high-spin ground states with the total number of valence d-electrons ranging from 2 to 4 with ground state multiplicities of 3 or 4 and excited state multiplicities of 1 or 2, respectively. In this work, the term "intermediate-spin" is not used because only two states are mainly considered, that is, one "low-spin" and one "high-spin. For example, a triplet might be referred to as "high-spin" relative to a "low-spin" singlet even though it is possible there could be an even higher spin state as in a quintet.

The same summary is provided for complexes with $e_g \rightarrow e_g$ spin-forbidden transitions in Table 1 (see Supporting Information Tables 2 and 5 for models of complexes and a

list of descriptors). In this case, there are 23 octahedral complexes all with $d^8$ Ni(II) centers. All transitions are from a high-spin triplet ground state to an excited state singlet.

Spin-forbidden transitions, $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$, are considered at the bottom of Table 1 (see Supporting Information Tables 3 and 6 for models of complexes and a list of descriptors). There are 20 complexes including 1 Co(II), 8 Co(III), 1 Fe(II), 7 Fe(III), and 3 Mn(II). The various molecular orbital diagrams are shown in Table 1. In this case, all cobalt complexes have low-spin singlet ground states, with the exception of cof6 which has a quartet ground state. A pair of fecn6 complexes serve as examples of other low-spin ground state complexes, $Fe(III)(CN)_6^{-3}$ as a doublet ground state and $Fe(II)(CN)_6^{-4}$ as a singlet ground state. $Fe(III)(en)_3^{+3}$ has a doublet ground state. The rest of the complexes have a sextet ground state.

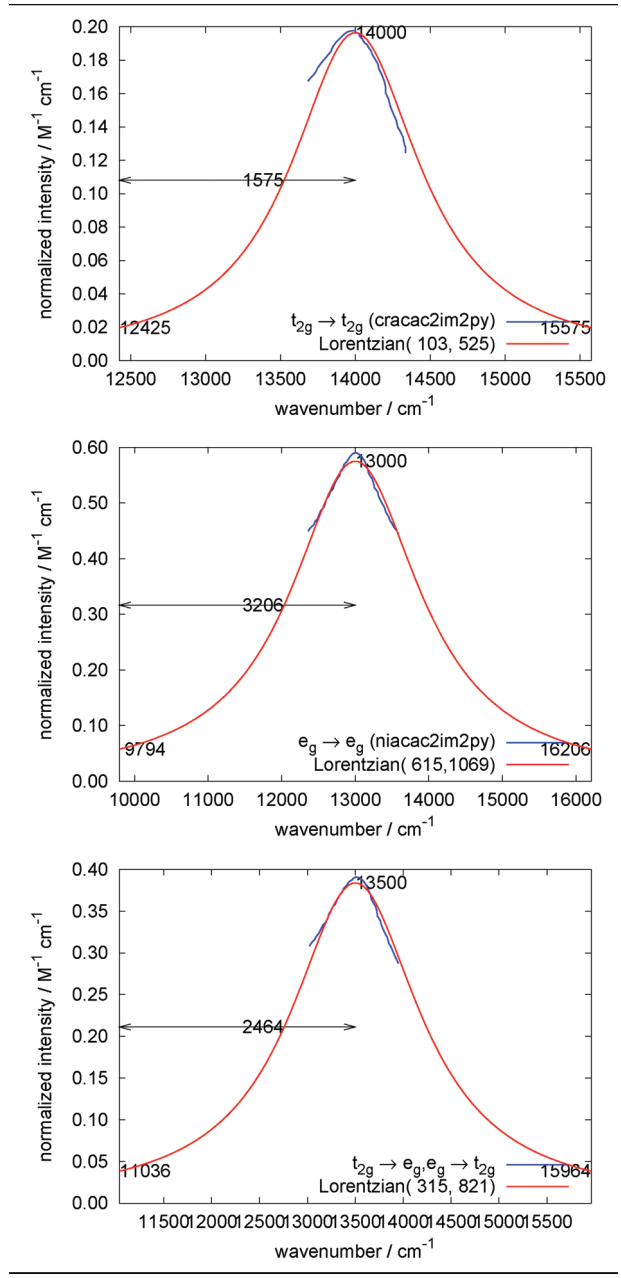
Minimized metal to coordinating atom bond distances are shown in Supporting Information Tables 7−9 for B3LYP/LACV3P minimizations in the appropriate environments for each of the spin states considered. During the minimizations, among a few other complexes, the excited state $d^8$ Ni(II) $O_h$ complexes with coordinating nitrogens tended to drop to $D_{4h}$ symmetry and become square planar when using B3LYP with a variety of basis sets. Most complexes are calculated to be slightly distorted octahedra where the metal−ligand bond distances are the degrees of freedom that undergo the largest change with a spin-forbidden transition. Bond distances are generally short in three particular cases (1) for states of lower multiplicity, (2) for complexes with a large number of valence electrons, and (3) for complexes with large metal oxidation states (see Supporting Information for a discussion).

**3.2. Estimates of Vibrational Relaxation Energies.** Estimates for the nonradiative vibrational relaxation energy for the $\nu_k \rightarrow \nu_0$ transition on the excited state surface are shown in Table 2. The difference in energy between the spin-forbidden highest intensity spectral peak $\nu_k$ and the lowest intensity spectral onset $\nu_0$ is determined by using digitizing software.[64] Spin-forbidden shoulder peaks are extracted from higher intensity MLCT and/or LMCT peaks of the experimental spectra and fit to a dilated Lorentzian with two free parameters $a$ and $b$

$$\mathcal{L}(x) = a\left(\frac{b}{(x - x_0)^2 + b^2}\right) \qquad (2)$$

where $x_0$ is the peak position. Relaxation energies for the experimental spectra of a number of different metal complexes with $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ spin-forbidden transitions were determined in this way. Using a unique vibrational relaxation energy for each complex was not possible, nor needed due to some inherent transferability in the electronic structure, because experimental spectra were not available in every case. Average peak to onset values of 1575 $cm^{-1}$ (4.50 ± 1.41 kcal/mol) and 3206 $cm^{-1}$ (9.17 ± 2.12 kcal/mol) were found. The results for cracac2im2py and niacac2im2py experiments, top and middle of Table 2, are representative examples of mean vibrational relaxation energies of $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ metal complexes and thus

***Table 2.*** Estimate of the Nonradiative Relaxation Energy from the Highest (Spectral Peak, $\nu_k$) to Lowest (Spectral Onset, $\nu_0$) Intensity Vibrational States for the Excited Electronic State[a]




[a] Vibrational relaxation energies of 1575 ($t_{2g} \rightarrow t_{2g}$ top), 3206 ($e_g \rightarrow e_g$ middle), and 2464 $cm^{-1}$ ($t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$ bottom) are obtained by subtracting the low intensity spectral onset (left) from the high intensity spectral peak. Spin-forbidden peaks, which typically show up as shoulder peaks to more intense MLCT or LMCT spin-conserving transitions, were resolved using digitizing software,[64] which extracts the coordinates from the experimental spectra shown in blue. These results were fit to a dilated Lorentzian, shown in red, with two free parameters.

are shown. More examples showing the fitted and experimental shoulder peaks for the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ spin-forbidden transitions along with estimates for the vibrational relaxation energies are shown in Table 10 of the Supporting Information. For each of the types of transitions the estimates of the vibrational relaxation energies are within 1−2 kcal/

mol of each other. The vibrational relaxation energy for the $e_g \rightarrow e_g$ complexes is larger than that for the $t_{2g} \rightarrow t_{2g}$ complexes because the later transition involves an excitation within the nonbonding $t_{2g}$ manifold, and so, there are less vibrational modes excited than for an excitation within the antibonding $e_g$ manifold. This is in agreement with the observation that the change in the average calculated metal to ligand bond distances accompanying excitation are greater for the $e_g \rightarrow e_g$ complexes than for the $t_{2g} \rightarrow t_{2g}$ complexes. Subtracting these values from the experimentally determined peak spin-forbidden transition energies allows direct comparison with calculated adiabatic transitions.

An estimate of the vibrational relaxation energy for the spin-forbidden $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$ transitions was determined using two different protocols. The first protocol involved fitting the shoulder peaks of the experiments for this class of complexes as was done for the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ transitions (see Table 2 and Supporting Information Table 10 for examples). The second protocol was to fit the vibrational relaxation energies of the $t_{2g} \rightarrow t_{2g}$ ($4.50 \pm 1.41$ kcal/mol) and $e_g \rightarrow e_g$ ($9.17 \pm 2.12$ kcal/mol) transitions determined above with their adiabatic changes in bond distances and interpolate the vibrational relaxation energy for the $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$ class of complexes. Using the average calculated metal−ligand bond distances, and one standard deviation to account for uncertainty, with values of 0.02 ($t_{2g} \rightarrow t_{2g}$), 0.15 ($e_g \rightarrow e_g$), and 0.12 Å ($t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$) (see the geometry section of the Supporting Information) an estimate for the vibrational relaxation energy was determined to be around 2800 $cm^{-1}$ which is between the values determined for $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ transitions. The fitting of the shoulder peaks also predicts the relaxation energy to lie between the relaxation energies of the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ transitions (see Supporting Information Table 10). Lastly an average of the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ vibrational relaxation energies provides a further comparison. The difference in estimates between the protocols is small, about 1 kcal/mol, considering the approximate 2 kcal/mol MUE we are aiming at in this work. The final estimate for the vibrational relaxation energy of the $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$ transitions is taken as $6.84 \pm 1.68$ kcal/mol. As previously mentioned this vibrational relaxation energy correction falls between the values determined for $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ transitions because here a single electron is moved to or from an antibonding $e_g$ orbital as opposed to moving an electron within nonbonding $t_{2g}$ or antibonding $e_g$ and so the effective response of the $e_g$ orbitals is intermediate. Further support for obtaining vibrational relaxation energies using this protocol is provided by noting that while the parameters of the DBLOC model were derived using vibrational relaxation corrected experimental gaps, those same DBLOC parameters provide very good results in comparison to experiments on near zero spin-crossover complexes for which the vibrational relaxation energy correction is zero.

**3.3. $t_{2g} \rightarrow t_{2g}$ Energetics.** A summary of experimental and calculated splittings in a dielectric continuum environment are shown in Tables 3 and 4 for the $t_{2g} \rightarrow t_{2g}$ spin-forbidden transitions. Experimental values, from which the 4.50 kcal/mol vibrational relaxation energy has been subtracted, range from 23.99 kcal/mol for vurea6 to 46.55 kcal/mol for mn2p2pameth2. With the exception of mn2p2pameth2 the experimental spin-forbidden transition energies cluster according to the molecular orbital diagrams at around 25 kcal/

***Table 3.*** Experimental and Calculated Spin-Forbidden Transition Energies (kcal/mol) in a Dielectric Continuum Environment[a]

| complex | mult. | $\langle S^2 \rangle$ | exp. | B3LYP | B3LYP error |
|---|---|---|---|---|---|
| cr223tetcl2 | 2 | 0.83 | 36.53 | 49.14 | −12.61 |
| cr223tetcl2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crcn6 | 2 | 0.76 | 31.00 | 47.47 | −16.47 |
| crcn6 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crcyclamncs2 | 2 | 0.79 | 34.54 | 49.51 | −14.97 |
| crcyclamncs2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| cren3 | 2 | 0.78 | 38.67 | 54.09 | −15.42 |
| cren3 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crf6 | 1 | 0.00 | 28.09 | 42.68 | −14.59 |
| crf6 | 3 | 2.02 | 0.00 | 0.00 | 0.00 |
| crnh34cl2 | 2 | 1.03 | 36.95 | 51.80 | −14.85 |
| crnh34cl2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crnh36 | 2 | 0.78 | 39.15 | 55.50 | −16.35 |
| crnh36 | 4 | 3.77 | 0.00 | 0.00 | 0.00 |
| crox3 | 2 | 0.86 | 36.53 | 48.83 | −12.31 |
| crox3 | 4 | 3.77 | 0.00 | 0.00 | 0.00 |
| mncn6 | 1 | 0.00 | 24.10 | 35.45 | −11.36 |
| mncn6 | 3 | 2.03 | 0.00 | 0.00 | 0.00 |
| mnf6 | 2 | 0.81 | 41.24 | 53.01 | −11.77 |
| mnf6 | 4 | 3.79 | 0.00 | 0.00 | 0.00 |

[a] For a pure spin ground state to a pure spin excited state, $t_{2g} \rightarrow t_{2g}$, the transition energy is larger than experiment.

***Table 4.*** Experimental and Calculated Spin-Forbidden Transition Energies (kcal/mol) in a Dielectric Continuum Environment[a]

| complex | mult. | $\langle S^2 \rangle$ | exp. | B3LYP | B3LYP error |
|---|---|---|---|---|---|
| cr223tetcl2 | 2 | 1.75 | 36.53 | 25.17 | 11.36 |
| cr223tetcl2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| cracac2im2py | 2 | 1.96 | 35.53 | 23.97 | 11.56 |
| cracac2im2py | 4 | 3.01 | 0.00 | 0.00 | 0.00 |
| crccsime36 | 2 | 1.76 | 33.52 | 22.95 | 10.57 |
| crccsime36 | 4 | 3.79 | 0.00 | 0.00 | 0.00 |
| crcn6 | 2 | 1.76 | 31.00 | 23.54 | 7.46 |
| crcn6 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crcyclamncs2 | 2 | 1.76 | 34.54 | 24.78 | 9.76 |
| crcyclamncs2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| cren3 | 2 | 1.75 | 38.67 | 27.04 | 11.63 |
| cren3 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crf6 | 2 | 1.75 | 40.37 | 25.89 | 14.48 |
| crf6 | 4 | 3.76 | 0.00 | 0.00 | 0.00 |
| crf6 | 1 | 1.00 | 28.09 | 13.13 | 14.96 |
| crf6 | 3 | 2.02 | 0.00 | 0.00 | 0.00 |
| crnh34cl2 | 2 | 1.75 | 36.95 | 26.08 | 10.87 |
| crnh34cl2 | 4 | 3.78 | 0.00 | 0.00 | 0.00 |
| crnh36 | 2 | 1.75 | 39.15 | 26.78 | 12.36 |
| crnh36 | 4 | 3.77 | 0.00 | 0.00 | 0.00 |
| crox3 | 2 | 1.75 | 36.53 | 24.58 | 11.95 |
| crox3 | 4 | 3.77 | 0.00 | 0.00 | 0.00 |
| mn2p2pameth2 | 1 | 0.00 | 46.55 | 39.94 | 6.62 |
| mn2p2pameth2 | 3 | 2.02 | 0.00 | 0.00 | 0.00 |
| mncn6 | 1 | 1.01 | 24.10 | 12.54 | 11.56 |
| mncn6 | 3 | 2.03 | 0.00 | 0.00 | 0.00 |
| mnf6 | 2 | 1.76 | 41.24 | 26.74 | 14.50 |
| mnf6 | 4 | 3.79 | 0.00 | 0.00 | 0.00 |
| vf6 | 1 | 1.00 | 24.66 | 12.53 | 12.13 |
| vf6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| vurea6 | 1 | 1.00 | 23.99 | 13.31 | 10.68 |
| vurea6 | 3 | 2.01 | 0.00 | 0.00 | 0.00 |

[a] For a pure spin ground state to a contaminated spin excited state, $t_{2g} \rightarrow t_{2g}$, the transition energy is smaller than experiment.

mol for $Cr(IV)F_6^{-2}$, mncn6, vf6, and vurea6, and at around 37 kcal/mol for the rest of the complexes. The DFT results cluster similarly. Note that the set of complexes that cluster around 25 kcal/mol are those that are ground state triplets, as opposed to ground state quartets like the other complexes, and so the lower spin-forbidden transition energy is attributed to a smaller spin−spin stabilization energy as can be seen in the molecular orbital diagrams in Table 1. In both tables, B3LYP predicts the correct ordering of ground and excited state multiplicities in every case. The effects of the environment were found to be significant for some of the highly charged complexes and in aligning the states, that is, reducing the standard deviation in the errors. Preliminary vacuum results for the chromium series, confining ourselves at the time to only the first couple of istate values, showed wildly, seemingly random, differing results for the spin-forbidden transition energies, by as much as 12 kcal/mol. With the use of solvent they were later found to correspond to different electronic states.

Table 3 shows that the B3LYP/LACV3P predicts one excited state to lie above the experiment by about 14 kcal/mol while in Table 4 it predicts a different excited state to lie below the experiment by about 11 kcal/mol. These are the two low-lying states present in the excited state manifold. Given that spin-forbidden transitions have zero dipole oscillator strength, some means of calculating the spin-forbidden transition intensity, Franck−Condon intensities[61] and potentially spin−orbit coupling terms[65] would be necessary to definitively tell which state corresponds to the correct mapping onto the shoulders observed in the experimental spectrum. Comparing Tables 3 and 4 shows that the high-energy state is a more restrictive spin pure state, while the low energy state is spin contaminated with the high-spin ground state. All ground states are pure spin with the exception of cracac2im2py. Three key facts suggest that the lower of the two spin states obtained from our B3LYP calculation is in fact the state corresponding to the shoulder seen in the optical absorption spectrum. First, as noted above, this state has ground state multiplicity mixed into the wave function, which facilitates the borrowing of intensity as compared to the higher energy state, which has no such admixture. Second, the higher energy state is not present in a significant number of our B3LYP computations, whereas the lower state is always present. This could be the result of failure to employ a suitable initial guess, but it also may be the case that the state is mixed with other excited states or has moved to a substantially higher energy. Third, the shoulder is the lowest-energy feature in the experimental optical absorption spectrum, and asserting its correspondence with the lowest-energy computed DFT state makes sense. Additionally, similar reasoning is used in a recent article by Watts et al.[63] where hybrid DFT methods are able to reproduce the experimental observation that temperature dependent magnetic moments of some iron-porphyrin complexes clearly show high-spin states for some ligands and a mix of high and intermediate spin states for others thereby resulting in a calculated pure spin state for the former and a mixed spin state for the later. On the basis of these arguments, we shall assume the stated correspondence in the analysis that follows. The achievement of consistency and accuracy in the numerical results for the energetics, if attained, will provide further evidence that this fundamental assumption that we are making is in fact correct.

Grouping the metal complexes in Table 1 according to molecular orbital diagrams and comparing to Table 4 shows that the average error for $t_{2g}^2$, $t_{2g}^3$, and $t_{2g}^4$ configurations is 12.59 kcal/mol, 11.50 kcal/mol, and 9.09 kcal/mol. B3LYP/LACV3P clearly has a large intrinsic error in the pairing of two $t_{2g}$ electrons and smaller errors for the spin−spin interactions. A preliminary systematic correction scheme in which $t_{2g}^2$ and $t_{2g}^4$ are treated as identical cases predicts a pairing correction of around 10 kcal/mol. The most striking characteristic of the data in Table 4 is that the errors, compared to experiment, are tightly grouped around the value of 10 kcal/mol, and hence a single empirical parameter of this magnitude can drastically reduce the error for complexes of the type $t_{2g} \rightarrow t_{2g}$. We interpret this value as the B3LYP/LACV3P error that is manifested when a d-electron in an organometallic complex is moved from a singly occupied $t_{2g}$ orbital to another singly occupied $t_{2g}$ orbital to form a doubly occupied $t_{2g}$ orbital. The sign of the correction implies that the low-spin form is overstabilized with respect to experiment. This is consistent with our previous work.[15,52] There, we observed that specific types of orbitals can manifest overbinding, due to overestimation of nondynamical correlation effects. The d-orbitals have a different shape and size, and further interact with ligands, hence the errors seen here are different from those in a bare metal atom.[15] We employ the same value, around 10 kcal/mol, of this correction for all metals in the $t_{2g} \rightarrow t_{2g}$ database and for all oxidation states. Whether the identical value can be used for the $e_g$ manifold is addressed in the next section. In our final empirical model, there are also corrections due to differing numbers of spin−spin interactions in the ground and excited states; this correction is relatively small, and is discussed further below when we present the final model which is optimized by least-squares fitting to all of the experimental data simultaneously.

**3.4. $e_g \rightarrow e_g$ Energetics.** Energies for the $e_g \rightarrow e_g$ spin-forbidden transitions of the Ni(II) complexes are shown in Table 5. Experimental triplet to singlet transitions, which have been reduced by 9.17 kcal/mol for the vibrational relaxation energy, range from 23.14 kcal/mol for nitach3mepyr to 33.72 kcal/mol for nif6. Table 5 shows the lower-energy spin-contaminated excited state, which has similar spin expectation values and errors in the energies as those values from Table 4. This is as expected because in both cases the molecular orbital diagrams in Table 1 show that the differences in the electronic structures between ground and excited states are simply the pairing of two electrons. The average error for the spin-forbidden transition energies is around 10 kcal/mol, a value consistent with the $t_{2g} \rightarrow t_{2g}$ transitions. This consistency validates both the protocol for analyzing the experimental data and the theoretical model explaining the origin of the DFT error in splittings for cases in which there is transfer of an electron within the $t_{2g}$ or $e_g$ manifolds, as opposed to between manifolds.

***Table 5.*** Experimental and Calculated Spin-Forbidden Transition Energies, $e_g \rightarrow e_g$, (kcal/mol) in a Dielectric Continuum Environment

| complex | mult. | $\langle S^2 \rangle$ | exp. | B3LYP | B3LYP error |
|---|---|---|---|---|---|
| ni12dimeim6 | 1 | 1.00 | 29.09 | 18.27 | 10.82 |
| ni12dimeim6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| ni1meim6 | 1 | 1.00 | 28.43 | 16.94 | 11.48 |
| ni1meim6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| ni2meim6 | 1 | 1.00 | 28.47 | 18.10 | 10.37 |
| ni2meim6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| niacac2im2py | 1 | 1.77 | 28.00 | 18.14 | 9.86 |
| niacac2im2py | 3 | 3.78 | 0.00 | 0.00 | 0.00 |
| nibipy3 | 1 | 0.00 | 25.11 | 13.07 | 12.04 |
| nibipy3 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nibpm2no3 | 1 | 0.92 | 27.64 | 15.76 | 11.88 |
| nibpm2no3 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nidmso6 | 1 | 1.00 | 30.94 | 17.46 | 13.48 |
| nidmso6 | 3 | 2.01 | 0.00 | 0.00 | 0.00 |
| nidpdpm2h2o2 | 1 | 1.00 | 28.95 | 17.66 | 11.29 |
| nidpdpm2h2o2 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nidpdpm2no3h2o | 1 | 0.99 | 28.95 | 17.22 | 11.74 |
| nidpdpm2no3h2o | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nidpdpmno32 | 1 | 0.10 | 28.95 | 14.02 | 14.94 |
| nidpdpmno32 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nidpdpmno32ch3cn | 1 | 0.89 | 28.95 | 17.22 | 11.73 |
| nidpdpmno32ch3cn | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| niedta | 1 | 1.00 | 27.00 | 16.84 | 10.16 |
| niedta | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nien2scn2 | 1 | 0.97 | 27.03 | 15.11 | 11.92 |
| nien2scn2 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nien3 | 1 | 0.71 | 26.08 | 8.09 | 17.99 |
| nien3 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nif6 | 1 | 1.00 | 33.72 | 19.41 | 14.31 |
| nif6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nigly3 | 1 | 1.00 | 28.29 | 17.42 | 10.87 |
| nigly3 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nih2o6 | 1 | 0.91 | 33.03 | 18.99 | 14.04 |
| nih2o6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| ninh36 | 1 | 1.00 | 28.23 | 17.35 | 10.87 |
| ninh36 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| niphen3 | 1 | 0.10 | 25.16 | 7.84 | 17.32 |
| niphen3 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nipyrazole6 | 1 | 1.00 | 28.86 | 17.52 | 11.34 |
| nipyrazole6 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |
| nitach3mepyr | 1 | 1.00 | 23.14 | 16.64 | 6.51 |
| nitach3mepyr | 3 | 2.01 | 0.00 | 0.00 | 0.00 |
| nitpm2 | 1 | 1.00 | 27.78 | 17.54 | 10.24 |
| nitpm2 | 3 | 2.01 | 0.00 | 0.00 | 0.00 |
| nitpmno32 | 1 | 0.91 | 27.58 | 17.87 | 9.71 |
| nitpmno32 | 3 | 2.00 | 0.00 | 0.00 | 0.00 |

***Table 6.*** Experimental and Calculated Spin-Forbidden Transition Energies, $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$, (kcal/mol), in a Dielectric Continuum Environment

| complex | mult. | $\langle S^2 \rangle$ | exp. | B3LYP | B3LYP error |
|---|---|---|---|---|---|
| coamn3s3sarh | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| coamn3s3sarh | 3 | 2.21 | 33.61 | 21.99 | 11.62 |
| coamn5ssarh | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| coamn5ssarh | 3 | 2.11 | 32.48 | 24.09 | 8.38 |
| coen3 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| coen3 | 3 | 2.04 | 32.33 | 29.32 | 3.01 |
| coetn4s2amp | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| coetn4s2amp | 3 | 2.13 | 32.62 | 25.27 | 7.35 |
| cof6 | 2 | 1.75 | 43.20 | 31.76 | 11.44 |
| cof6 | 4 | 3.75 | 0.00 | 0.00 | 0.00 |
| col2 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| col2 | 3 | 2.04 | 23.74 | 21.97 | 1.77 |
| con3s3 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| con3s3 | 3 | 2.18 | 34.72 | 22.36 | 12.36 |
| conh35soch32 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| conh35soch32 | 3 | 2.04 | 26.22 | 23.42 | 2.80 |
| conh36 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| conh36 | 3 | 2.04 | 30.33 | 29.79 | 0.55 |
| fecat3 | 4 | 3.79 | 22.61 | 22.59 | 0.02 |
| fecat3 | 6 | 8.76 | 0.00 | 0.00 | 0.00 |
| fecn6 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| fecn6 | 3 | 2.04 | 60.93 | 37.40 | 23.52 |
| fecn6 | 2 | 0.76 | 0.00 | 0.00 | 0.00 |
| fecn6 | 4 | 3.80 | 45.14 | 28.58 | 16.57 |
| feen3 | 2 | 0.79 | 0.00 | 0.00 | 0.00 |
| feen3 | 4 | 3.81 | 18.89 | 20.91 | −2.02 |
| feeta3 | 4 | 3.79 | 25.19 | 21.87 | 3.32 |
| feeta3 | 6 | 8.76 | 0.00 | 0.00 | 0.00 |
| feh2o6 | 4 | 3.77 | 29.19 | 23.84 | 5.35 |
| feh2o6 | 6 | 8.76 | 0.00 | 0.00 | 0.00 |
| fethiocarbamate3 | 4 | 3.99 | 11.75 | 6.57 | 5.18 |
| fethiocarbamate3 | 6 | 8.76 | 0.00 | 0.00 | 0.00 |
| fetrencam | 4 | 3.80 | 22.90 | 24.21 | −1.31 |
| fetrencam | 6 | 8.76 | 0.00 | 0.00 | 0.00 |
| mnden2 | 4 | 3.77 | 39.05 | 28.64 | 10.42 |
| mnden2 | 6 | 8.75 | 0.00 | 0.00 | 0.00 |
| mnen3 | 4 | 3.77 | 38.05 | 27.71 | 10.34 |
| mnen3 | 6 | 8.75 | 0.00 | 0.00 | 0.00 |
| mnh2o6 | 4 | 3.76 | 47.12 | 43.44 | 3.67 |
| mnh2o6 | 6 | 8.75 | 0.00 | 0.00 | 0.00 |

**3.5. $t_{2g} \rightarrow e_g$ and $e_g \rightarrow t_{2g}$ Energetics.** There is much greater variation in the electronic structures shown in Table 1 for the last set of spin-forbidden transitions because there are both excitations, $t_{2g} \rightarrow e_g$, and de-excitations, $e_g \rightarrow t_{2g}$, on going from a ground state of one multiplicity to an excited state of another multiplicity. The $t_{2g} \rightarrow e_g$ and $e_g \rightarrow t_{2g}$ database has the largest variation in experimental results with transition energies ranging from 11.75 kcal/mol for fethiocarbamate3 to 60.93 kcal/mol for Fe(II)$(CN)_6^{-4}$. Please see the Supporting Information discussion section for more details about the experimental results. Here all of the cobalt complexes, with the exception of cof6, along with the two fecn6 complexes (Fe(II)$(CN)_6^{-4}$ and Fe(III)$(CN)_6^{-3}$) and the feen3 complex, have low-spin ground states while the rest of the complexes have high-spin ground states like the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ complexes. Any set of empirical corrections to the B3LYP results must be robust enough to ensure both consistency and accuracy across this database, a task which may be challenging considering the substantial variation observed in the B3LYP error in Table 6. Emphasis is on comparing the B3LYP with the experimental spin-forbidden transition energies keeping in mind the ligand field splitting parameter, $\Delta_o$, that empirically depends on ligands according to the spectrochemical series as well as on metal oxidation state.

While an electron changes manifolds, $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$, the cases being considered here still involve interconversion between two singly occupied orbitals and one doubly occupied orbital, although the singly occupied orbitals are now in two different manifolds. Thus, the pairing correction described previously will still be enforced, as will the spin−spin interaction correction. However, a major new possible source of error has been introduced, which is the DFT calculation of the splitting between the $t_{2g}$ and $e_g$ manifolds. From the data in Table 6, it is clear that this error must be significant, as the error pattern across the various complexes is quite different from the uniformity seen in the previous two classes of complexes, exhibiting substantial variation in magnitude.

A review of the basic physics of the origin of the splitting provides a useful starting point for understanding how the DFT error is likely to depend upon the chemical nature of the transition metal complex. The $t_{2g}$ orbitals have minimal interaction with the ligand lone pairs that are pointed at the metal, and are to a first approximation viewed as nonbonding orbitals. The $e_g$ orbitals in contrast can have a strong interaction with the ligand orbitals. The importance of this interaction depends primarily upon two factors: the energy of the relevant ligand orbitals, and overlap of these orbitals with the $e_g$ metal d-orbitals. Qualitatively, higher energies of the ligand orbitals (bringing them closer in energy to the metal d-orbitals) and greater overlap of ligand and metal d-orbitals leads to a larger mixing of the ligand and $e_g$ orbitals, and this in turn is reflected in an increase in the energy of the $e_g$ orbitals which essentially take on the role of low lying antibonding orbitals of the complex. The ligand orbitals that interact with the $e_g$ orbitals in turn play the role of bonding orbitals, and are pushed downward in energy by greater mixing. When the $t_{2g} \rightarrow e_g$ splitting is large compared to the pairing energy of two d-electrons, the $t_{2g}$ orbitals are filled prior to any occupation of the $e_g$ orbitals, leading to a low-spin complex in the ground state. In contrast, when the splitting is small compared to the pairing energy, the $e_g$ orbitals are singly occupied prior to double occupation of any $t_{2g}$ orbitals, and a high-spin state becomes the ground state.

DFT based errors in the $\Delta_o$ splitting must be dominated by errors in the matrix elements between the ligand and metal d-orbitals that interact to form the new $e_g$ orbitals. The metal−ligand "bonding" characterizing such interactions can be viewed as having a large ionic character. There is a strong dependence of these nondynamical correlation errors on DFT functional, which explains why different functionals yield very different prediction of relative transition metal spin state energetics. For B3LYP, in an ionic complex such as NaCl, the strength of the ionic bond (which in turn is proportional to the Hamiltonian matrix element between the appropriate $Na^+$ and $Cl^-$ orbitals) is underestimated by 4.5 kcal/mol.[52] Such an underestimation will in turn lead to an equivalent underestimation of the splitting of the $t_{2g}$ and $e_g$ orbitals, and this underestimation will stabilize high-spin, as opposed to low-spin, states. This is the origin of the observation in the literature concerning the apparent overstabilization of high-spin complexes by B3LYP and other hybrid functionals.

We expect each ligand lone pair coordinated to the metal to contribute a component to the error. The magnitude of the contribution should be on the same energy scale as the bond energy errors seen for organic systems, roughly 1−5 kcal/mol. Finally, the value of the error will depend predominantly upon the ligand and, secondarily, upon the metal and oxidation state, and we expect to see a semiquantitative correlation with the spectrochemical series. The most accurate model would be derived by assigning a large number of parameters for different ligands depending upon their chemical structure, however, we lack sufficient data to do this in a fine grained fashion. Hence, to avoid overfitting, we instead try to use the smallest number of ligand parameters possible in the analysis that follows. This leads to a "minimalist" model which nevertheless displays remarkable predictive power. Future versions of the model aimed at higher accuracy will likely need to increase the number of ligand parameters as is briefly discussed below.

By grouping the complexes in Table 1 according to their molecular orbital diagrams, as well as coordinating atoms and metal oxidation states, it can be seen that the experimental results cluster accordingly and are consistent considering the spectrochemical series and spin−spin stabilization energies. Again it is seen that a balance in the electronic structure becomes important for example consider the similarity in the spin-splitting energies for complexes with differing electronic structures as in Co complexes with $t_{2g} \rightarrow e_g$ versus Fe complexes with $e_g \rightarrow t_{2g}$, as well as balancing the effect of decreasing $\Delta_o$ with an oxidation state less than two with increasing $\Delta_o$ with strongly interacting ligands as in $Fe(II)(CN)_6^{-4}$. In every case the correct ordering of the states is obtained with B3LYP and in many cases the conventional B3LYP was already close to experiment.

Each Co(III)-N6 complex, coen3, col2, conh35soch32, and conh36 involves an unpairing of a $t_{2g}$ orbital worth around −10 kcal/mol. This is the same pairing error as determined before for the $t_{2g} \rightarrow t_{2g}$ and $e_g \rightarrow e_g$ complexes but with opposite sign. By comparing the −10 kcal/mol with the average error in the spin-forbidden transition energy, which is already reasonably small, around 2 kcal/mol, a $\Delta_o$ correction is determined to be around 2 kcal/mol per coordinating nitrogen. As will be discussed below all values will be optimized by linear regression for the final DBLOC model. Proceeding similarly for the Co(III) nitrogen and sulfur complexes, coamn3s3sarh, coamn5ssarh, coetn4s2amp, and con3s3, there is an unpairing correction of −10 kcal/mol along with the 2 kcal/mol per nitrogen for the $t_{2g} \rightarrow e_g$ excitation. Comparison to the experimental gap provides a larger $\Delta_o$ correction of 5 kcal/mol per sulfur atom. Note that with the exception of the small outlier coamn5ssarh the error increases as the number of coordinating thio-ether sulfurs increases.

For higher multiplicity calculations the singly occupied−singly occupied ($t_{2g}$ or $e_g$) interaction becomes important and depending on the oxidation state and ligands very small $\Delta_o$ can induce some interaction between $t_{2g}$ and $e_g$ electrons. Considering fethiocarbamate3 there are corrections of 10 kcal/mol for pairing and a total of −12 kcal/mol for de-exciting in a field of six anionic sulfurs. The later comes from the fact that anionic sulfurs lie to the left of the spectrochemical series and thus receive a 2 kcal/mol correction per atom. Comparing the sum with the error in the spin-forbidden transition of 5 kcal/mol gives an estimate for creating a singly occupied−singly occupied interaction at around −1 kcal/mol. The fethiocarbamate3 despite being a +3 oxidation state complex loses 7 such interactions because anionic sulfurs are so far left in the spectrochemical series thereby inducing spin−spin stabilization between the $t_{2g}$ and $e_g$ manifolds. The Fe(III)−O6 complexes having a small average error can be analyzed using the same types of corrections as for fethiocarbamate3, except that in this case,

***Table 7.*** Final Five DBLOC Correction Parameters for Spin-Forbidden Transitions Optimized by Linear Regression[a]

| symbol | description | value (kcal/mol) |
|---|---|---|
| p | create a doubly occupied orbital | 10.05/pair |
| ss | create a singly occupied−singly occupied interaction | −1.05/interaction |
| exlss | $t_{2g} \rightarrow e_g$ excitation (left spectrochemical series) | 1.88/atom |
| exmss | $t_{2g} \rightarrow e_g$ excitation (middle spectrochemical series) | 2.85/atom |
| exrss | $t_{2g} \rightarrow e_g$ excitation (right spectrochemical series) | 5.21/atom |

[a] The values shown are for the pairing of electrons, the creation of a spin−spin interaction, or for $t_{2g} \rightarrow e_g$ excitation for the excited state relative to the ground state. Note that the last three parameters are corrections for $\Delta_o$.

we only lose 3 spin−spin interactions and the de-excitation is in a field of six oxygens. This includes both neutral oxygen, as in the aquo ligand, and anionic oxygen, as in the hydroxyl ligand. These ligands are again to the left of the spectrochemical series and so the same parameters are used giving reasonable results. For the Mn(II)−N6 and Mn(II)−O6 complexes, there are pairing and de-excitation parameters along with the loss of 7 spin−spin interactions because of the +2 oxidation state. Co(II)-F6 has a total correction of around 11 kcal/mol which is composed only of the pairing correction and the loss of 3 spin−spin corrections. This is because the oxidation state is +2 and the fluoride ion lies far to the left of the spectrochemical series. This complex has roughly 5-fold degenerate d-orbitals and is more comparable to the $t_{2g} \rightarrow t_{2g}$ or $e_g \rightarrow e_g$ class of complexes. Since the cyano ligand is to the far right of the spectrochemical series it is more comparable to the 5 kcal/mol correction for the thio-ether sulfur containing ligands, and so Fe(II)−C6, along with the pairing correction, results in a total correction of 20 kcal/mol. Analyzing the Fe(III)−C6 complex proceeds similarly except for the formation of a spin−spin interaction.

**3.6. DBLOC Model.** A simplified correction scheme can now be devised for calculating accurate relative spin state energetics of transition metal complexes. An acceptable ratio of parameters to data points has been achieved. The final set of five DBLOC parameters optimized by linear regression are shown in Table 7 and the errors with respect to experiment of the corrected spin-forbidden transition energies are shown in Table 8. The parameter exmss will be discussed in the section below on small-gap spin-crossover complexes but is included now for completeness. The DBLOC energy, $E_{\mathrm{DBLOC}}$, is given by adding the DBLOC correction, $E_{\mathrm{DBLOC}}^{\mathrm{corr}}$, to the B3LYP energy, $E_{\mathrm{B3LYP}}$,

$$E_{\mathrm{DBLOC}} = E_{\mathrm{B3LYP}} + E_{\mathrm{DBLOC}}^{\mathrm{corr}} \quad (3)$$

where

***Table 8.*** Errors (kcal/mol) with Respect to Experiment for the DBLOC Corrected Spin-Forbidden Transition Energies, along with the Weights of the Parameters Shown in Table 7[a]

| complex | p | ss | exlss | exmss | exrss | exp. | B3LYP error | DBLOC error |
|---|---|---|---|---|---|---|---|---|
| cr223tetcl2 | 1 | −3 | 0 | 0 | 0 | 36.53 | 11.36 | −1.82 |
| cracac2im2py | 1 | −3 | 0 | 0 | 0 | 35.53 | 11.56 | −1.63 |
| crccsime36 | 1 | −3 | 0 | 0 | 0 | 33.52 | 10.57 | −2.61 |
| crcn6 | 1 | −3 | 0 | 0 | 0 | 31.00 | 7.46 | −5.73 |
| crcyclamncs2 | 1 | −3 | 0 | 0 | 0 | 34.54 | 9.76 | −3.43 |
| cren3 | 1 | −3 | 0 | 0 | 0 | 38.67 | 11.63 | −1.56 |
| crf6 | 1 | −3 | 0 | 0 | 0 | 40.37 | 14.48 | 1.30 |
| crf6 | 1 | −1 | 0 | 0 | 0 | 28.09 | 14.96 | 3.87 |
| crnh34cl2 | 1 | −3 | 0 | 0 | 0 | 36.95 | 10.87 | −2.31 |
| crnh36 | 1 | −3 | 0 | 0 | 0 | 39.15 | 12.36 | −0.82 |
| crox3 | 1 | −3 | 0 | 0 | 0 | 36.53 | 11.95 | −1.24 |
| mn2p2pameth2 | 1 | −1 | 0 | 0 | 0 | 46.55 | 6.62 | −4.47 |
| mncn6 | 1 | −1 | 0 | 0 | 0 | 24.10 | 11.56 | 0.47 |
| mnf6 | 1 | −3 | 0 | 0 | 0 | 41.24 | 14.50 | 1.31 |
| vf6 | 1 | −1 | 0 | 0 | 0 | 24.66 | 12.13 | 1.04 |
| vurea6 | 1 | −1 | 0 | 0 | 0 | 23.99 | 10.68 | −0.41 |
| ni12dimeim6 | 1 | −1 | 0 | 0 | 0 | 29.09 | 10.82 | −0.27 |
| ni1meim6 | 1 | −1 | 0 | 0 | 0 | 28.43 | 11.48 | 0.39 |
| ni2meim6 | 1 | −1 | 0 | 0 | 0 | 28.47 | 10.37 | −0.72 |
| niacac2im2py | 1 | −1 | 0 | 0 | 0 | 28.00 | 9.86 | −1.23 |
| nibipy3 | 1 | −1 | 0 | 0 | 0 | 25.11 | 12.04 | 0.95 |
| nibpm2no3 | 1 | −1 | 0 | 0 | 0 | 27.64 | 11.88 | 0.79 |
| nidmso6 | 1 | −1 | 0 | 0 | 0 | 30.94 | 13.48 | 2.39 |
| nidpdpm2h2o2 | 1 | −1 | 0 | 0 | 0 | 28.95 | 11.29 | 0.20 |
| nidpdpm2no3h2o | 1 | −1 | 0 | 0 | 0 | 28.95 | 11.74 | 0.64 |
| nidpdpmno32 | 1 | −1 | 0 | 0 | 0 | 28.95 | 14.94 | 3.84 |
| nidpdpmno32ch3cn | 1 | −1 | 0 | 0 | 0 | 28.95 | 11.73 | 0.64 |
| niedta | 1 | −1 | 0 | 0 | 0 | 27.00 | 10.16 | −0.93 |
| nien2scn2 | 1 | −1 | 0 | 0 | 0 | 27.03 | 11.92 | 0.83 |
| nien3 | 1 | −1 | 0 | 0 | 0 | 26.08 | 17.99 | 6.89 |
| nif6 | 1 | −1 | 0 | 0 | 0 | 33.72 | 14.31 | 3.21 |
| nigly3 | 1 | −1 | 0 | 0 | 0 | 28.29 | 10.87 | −0.22 |
| nih2o6 | 1 | −1 | 0 | 0 | 0 | 33.03 | 14.04 | 2.95 |
| ninh36 | 1 | −1 | 0 | 0 | 0 | 28.23 | 10.87 | −0.22 |
| niphen3 | 1 | −1 | 0 | 0 | 0 | 25.16 | 17.32 | 6.23 |
| nipyrazole6 | 1 | −1 | 0 | 0 | 0 | 28.86 | 11.34 | 0.25 |
| nitach3mepyr | 1 | −1 | 0 | 0 | 0 | 23.14 | 6.51 | −4.59 |
| nitpm2 | 1 | −1 | 0 | 0 | 0 | 27.78 | 10.24 | −0.85 |
| nitpmno32 | 1 | −1 | 0 | 0 | 0 | 27.58 | 9.71 | −1.38 |
| coamn3s3sarh | −1 | 0 | 3 | 0 | 3 | 33.61 | 11.62 | 0.42 |
| coamn5ssarh | −1 | 0 | 5 | 0 | 1 | 32.48 | 8.38 | 3.85 |
| coen3 | −1 | 0 | 6 | 0 | 0 | 32.33 | 3.01 | 1.81 |
| coetn4s2amp | −1 | 0 | 4 | 0 | 2 | 32.62 | 7.35 | −0.52 |
| cof6 | 1 | −3 | 0 | 0 | 0 | 43.20 | 11.44 | −1.75 |
| col2 | −1 | 0 | 6 | 0 | 0 | 23.74 | 1.77 | 0.56 |
| con3s3 | −1 | 0 | 3 | 0 | 3 | 34.72 | 12.36 | 1.16 |
| conh35soch32 | −1 | 0 | 6 | 0 | 0 | 26.22 | 2.80 | 1.59 |
| conh36 | −1 | 0 | 6 | 0 | 0 | 30.33 | 0.55 | −0.66 |
| fecat3 | 1 | −3 | −6 | 0 | 0 | 22.61 | 0.02 | −1.91 |
| fecn6 | −1 | 0 | 0 | 0 | 6 | 60.93 | 23.52 | 2.33 |
| fecn6 | −1 | 1 | 0 | 0 | 6 | 45.14 | 16.57 | −3.58 |
| feen3 | −1 | 1 | 6 | 0 | 0 | 18.89 | −2.02 | −2.18 |
| feeta3 | 1 | −3 | −6 | 0 | 0 | 25.19 | 3.32 | 1.38 |
| feh2o6 | 1 | −3 | −6 | 0 | 0 | 29.19 | 5.35 | 3.42 |
| fethiocarbamate3 | 1 | −7 | −6 | 0 | 0 | 11.75 | 5.18 | −0.94 |
| fetrencam | 1 | −3 | −6 | 0 | 0 | 22.90 | −1.31 | −3.24 |
| mnden2 | 1 | −7 | −6 | 0 | 0 | 39.05 | 10.42 | 4.30 |
| mnen3 | 1 | −7 | −6 | 0 | 0 | 38.05 | 10.34 | 4.22 |
| mnh2o6 | 1 | −7 | −6 | 0 | 0 | 47.12 | 3.67 | −2.45 |

[a] The $t_{2g} \rightarrow t_{2g}$, $e_g \rightarrow e_g$, and $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$ are at the top, middle, and bottom, respectively.

$$E_{\mathrm{DBLOC}}^{\mathrm{corr}} = \frac{1}{2}({}^{0}n_{\alpha}^{t_{2g}} - {}^{k}n_{\alpha}^{t_{2g}} + {}^{0}n_{\alpha}^{e_g} - {}^{k}n_{\alpha}^{e_g})p + \lambda({}^{k}n_{\alpha}^{t_{2g}}{}^{k}n_{\alpha}^{e_g} - {}^{0}n_{\alpha}^{t_{2g}}{}^{0}n_{\alpha}^{e_g})\mathrm{ss} + \frac{1}{2}({}^{k}n_{\alpha}^{t_{2g}}({}^{k}n_{\alpha}^{t_{2g}} - 1) + {}^{k}n_{\alpha}^{e_g}({}^{k}n_{\alpha}^{e_g} - 1) - {}^{0}n_{\alpha}^{t_{2g}}({}^{0}n_{\alpha}^{t_{2g}} - 1) - {}^{0}n_{\alpha}^{e_g}({}^{0}n_{\alpha}^{e_g} - 1))\mathrm{ss} + ({}^{0}n^{t_{2g}} - {}^{k}n^{t_{2g}}) \times \sum_{i=1}^{6}(\mathrm{exlss}\delta_{L_i 0} + \mathrm{exmss}\delta_{L_i 1} + \mathrm{exrss}\delta_{L_i 2}) \quad (4)$$

$$\lambda \approx H(2 - q)H(3 - (L_1 + L_2 + L_3 + L_4 + L_5 + L_6)) \quad (5)$$

with ${}^{0}n_{\alpha}^{t_{2g}}$, ${}^{k}n_{\alpha}^{t_{2g}}$, ${}^{0}n_{\alpha}^{e_g}$, and ${}^{k}n_{\alpha}^{e_g}$ being the number of unpaired $t_{2g}$ or $e_g$ electrons for the ground, 0th, and excited states,

$k$th, respectively. The total number of $t_{2g}$ electrons for the ground and excited states are given by ${}^{0}n^{t_{2g}}$ and ${}^{k}n^{t_{2g}}$. The parameter $\lambda$ turns on spin−spin interactions between the $t_{2g}$ and $e_g$ manifolds for the case of small $\Delta_o$, that is, where the Heaviside functions, $H(2 - q)$ or $H(3 - (L_1 + L_2 + L_3 + L_4 + L_5 + L_6))$, become positive in cases where the metal has oxidation state less than or equal to 2 or when the sum of ligand integers (discussed below) is less than or equal to around 3. The sum is over the first coordination sphere of an octahedral complex where the value of $L_i$ indicates wether the coordinating atom lies to the left ($L_i = 0$), middle ($L_i = 1$), or right ($L_i = 2$) of the spectrochemical series and the Kronecker delta selects the appropriate term. Parameters p, ss, exlss, exmss, and exrss are the DBLOC model parameters. Sign conventions for the parameters are as shown for either creating an electron pair (p), creating a spin−spin interaction (ss), or for moving an electron from $t_{2g} \rightarrow e_g$ (exlss, exmss, and exrss, depending upon the ligand atoms coordinated to the metal) as the complex transitions from the ground to the excited state. Note that the parameter ss is applied per interaction, while the parameters exlss, exmss, and exrss are applied per coordinating atom. Opposite signed parameters must be used if the ground to excited state transition involves the opposite of the listed physical process. This reversibility of the parameters with reversibility of the physical interpretation provides support for the robustness of the underlying model. For example, the pairing parameter at 10.05 kcal/mol changes sign if unpairing electrons. The magnitude of this value signifies the importance of errors in B3LYP's description of nondynamical correlation of occupied d-orbitals in metal complexes. The $t_{2g} \rightarrow e_g$ excitation parameters, exrss, exmss, and exlss, are different because the error in the B3LYP depends on the position of the coordinating atoms in the spectrochemical series. For this database the error increases from 1.88 to 2.85 and finally to 5.21 kcal/mol as we move to the right of the spectrochemical series corresponding to a larger $\Delta_o$ and thus more strongly interacting ligands. As noted above, the signs and magnitudes of these corrections, when interpreted in a simple ligand field picture, are in remarkably good correspondence with the analogous range of correction parameters analyzed for metal atoms[15] and organic molecules.[52]

In general, B3LYP largely underestimates the energy required to pair two $t_{2g}$ or $e_g$ electrons, overestimates the energy to unpair them, slightly overestimates the energy of a parallel spin−spin interaction, and underestimates the size of $\Delta_o$, more severely for strongly interacting ligands. Table 8 shows that the MUE for conventional B3LYP goes from 11.40 kcal/mol (standard deviation of 2.25 kcal/mol) to 2.13 kcal/mol (standard deviation of 1.53 kcal/mol) for $t_{2g} \rightarrow t_{2g}$, from 11.95 kcal/mol (standard deviation of 2.51 kcal/mol) to 1.77 kcal/mol (standard deviation of 1.96 kcal/mol) for $e_g \rightarrow e_g$, and from 7.05 kcal/mol (standard deviation of 6.06 kcal/mol) to 2.11 kcal/mol (standard deviation of 1.27 kcal/mol) for $t_{2g} \rightarrow e_g$ or $e_g \rightarrow t_{2g}$. This is an overall decrease from 10.14 kcal/mol (standard deviation of 4.56 kcal/mol and maximum of 23.52 kcal/mol) to 1.98 kcal/mol (standard deviation of 1.62 kcal/mol and maximum of 6.89 kcal/mol). This maximum error is for nien3 which is one of the complexes for which B3LYP/LACV3P has some difficulties in breaking the octahedral symmetry. To summarize, the DBLOC model performs extremely well energetically and removes many outliers. Complexes crcn6, mn2p2pameth2, nitach3mepyr, and feh2o6 show only minor improvement while the error for some complexes increases only slightly.

***Table 9.*** Comparison of Errors for Conventional B3LYP (20% Exact Non-Local Exchange), B3LYP* (15% Exact Non-Local Exchange), and DBLOC with Experimental Spin-Forbidden Transition Energies (kcal/mol) for Selected Complexes, crnh36 ($t_{2g} \rightarrow t_{2g}$), nigly3 ($e_g \rightarrow e_g$), cof6 and fetrencam ($e_g \rightarrow t_{2g}$), and fecn6 ($t_{2g} \rightarrow e_g$)[a]

| complex | mult. | exp. | B3LYP error | B3LYP* error | error diff. | DBLOC error |
|---|---|---|---|---|---|---|
| crnh36 | 2 | 39.15 | 12.36 | 15.06 | 2.70 | −0.83 |
| crnh36 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| nigly3 | 1 | 28.29 | 10.85 | 16.15 | 5.31 | −0.24 |
| nigly3 | 3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| cof6 | 2 | 43.20 | 11.45 | 19.75 | 8.30 | −1.74 |
| cof6 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| fecn6 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| fecn6 | 4 | 45.14 | 16.59 | 12.82 | −3.77 | −3.56 |
| fetrencam | 4 | 22.90 | −1.32 | 0.52 | 1.84 | −3.25 |
| fetrencam | 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

[a] For all complexes but fecn6, the ground state is high-spin. B3LYP's stabilization of the high-spin states relative to B3LYP* results in better agreement with experiment in all cases except fecn6. The DBLOC model shows considerable improvement over both B3LYP and B3LYP*. The column error diff is the difference between B3LYP* and B3LYP errors.

Application of the DBLOC model to 7 metal complexes studied by Swart et al.[66] (see Supporting Information Tables 11−13) shows systematic results for similar chemical species and illustrates that although it is possible for the model to reorder the spin states, the correct ordering of the states given by conventional DFT is preserved by DBLOC. The spin multiplicities of the ground states of these complexes are known, however, the experimental gaps remain unknown.

**3.7. Comparison to B3LYP*.** Comparison of errors for conventional B3LYP, B3LYP* (5% less exact nonlocal exchange than B3LYP), and DBLOC single point LACV3P calculations with experiment for some selected complexes are shown in Table 9. The complexes are crnh36, which is $t_{2g} \rightarrow t_{2g}$, nigly3, which is $e_g \rightarrow e_g$, cof6 and fetrencam, which are $e_g \rightarrow t_{2g}$, and fecn6, which is $t_{2g} \rightarrow e_g$, and experimentally all are high-spin ground states with the exception of fecn6. The results shown for B3LYP and DBLOC and their comparison with experiment are the same as the results previously discussed. As expected, comparing the B3LYP* and B3LYP total energies (not shown) shows that reducing the amount of exact nonlocal exchange makes the energies less negative. The important differences, shown in the relative energies, are that the high-spin states undergo a larger change than the low-spin states resulting in smaller spin gaps for complexes with high-spin ground states and larger spin gaps for complexes with low-spin ground states. For these experimental spin-forbidden transitions reducing the amount of exact nonlocal exchange brings the final B3LYP* result further from experiment by about 5 kcal/mol on average in every case except fecn6 for which the results are closer to experiment by about 4 kcal/mol. Similar problems would be manifested in many other complexes in our database. In

***Table 10.*** Molecular Orbital Diagrams for the Octahedral Small-Gap Spin-Crossover Complexes Studied in Refs 47, 48, and 67

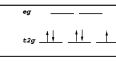
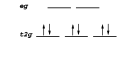
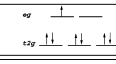
| complex | type | num. val. elec. | g.s. mult. | g.s. l.f.d. | e.s. mult. | e.s. l.f.d. |
|---|---|---|---|---|---|---|
| feacac2trien | $t_{2g} \rightarrow e_g$ | 5 | 2 |  | 6 |  |
| fepapth2, fetacn2, fe2amp3, fehbpz32, fepybzimh3, fetppn3, fephen2ncs2, fetpen, fetpancs2, febptnncs2, fephen2ncse2 | $t_{2g} \rightarrow e_g$ | 6 | 1 |  | 5 |  |
| coterpy2, copyimine22 | $t_{2g} \rightarrow e_g$ | 7 | 2 |  | 4 |  |

***Table 11.*** Application of DBLOC to Spin-Crossover Complexes with Near-Zero Spin Gap[a]

| complex | p | ss | exlss | exmss | exrss | exp. | B3LYP | B3LYP error | DBLOC | DBLOC error |
|---|---|---|---|---|---|---|---|---|---|---|
| feacac2trien | −2 | 4 | 8 | 4 | 0 | 2.87 | −3.11 | 5.98 | −0.97 | 3.84 |
| fepapth2 | −2 | 2 | 0 | 12 | 0 | 3.82 | −8.13 | 11.95 | 3.87 | −0.05 |
| fetacn2 | −2 | 2 | 12 | 0 | 0 | 5.50 | −5.02 | 10.52 | −4.66 | 10.16 |
| fe2amp3 | −2 | 2 | 6 | 6 | 0 | 5.26 | −7.17 | 12.43 | −0.99 | 6.25 |
| fehbpz32 | −2 | 2 | 0 | 12 | 0 | 4.54 | −0.72 | 5.26 | 11.28 | −6.74 |
| fepybzimh3 | −2 | 2 | 0 | 12 | 0 | 5.02 | −6.93 | 11.95 | 5.07 | −0.05 |
| fetppn3 | −2 | 2 | 4 | 8 | 0 | 6.69 | −5.26 | 11.95 | 2.86 | 3.83 |
| coterpy2 | −1 | 1 | 0 | 6 | 0 | 3.11 | −4.78 | 7.89 | 1.22 | 1.89 |
| copyimine22 | −1 | 1 | 0 | 6 | 0 | 3.35 | −6.21 | 9.56 | −0.21 | 3.56 |
| fephen2ncs2 | −2 | 2 | 4 | 8 | 0 | 0.00 | −7.98 | 7.98 | 0.14 | −0.14 |
| fetpen | −2 | 2 | 4 | 8 | 0 | 0.00 | −2.29 | 2.29 | 5.83 | −5.83 |
| fetpancs2 | −2 | 2 | 6 | 6 | 0 | 0.00 | −3.99 | 3.99 | 2.19 | −2.19 |
| febptnncs2 | −2 | 2 | 8 | 4 | 0 | 0.00 | −8.22 | 8.22 | −3.98 | 3.98 |
| fephen2ncse2 | −2 | 2 | 4 | 8 | 0 | 0.00 | −7.43 | 7.43 | 0.69 | −0.69 |

[a] B3LYP and experimental results are taken from the literature. The first nine complexes are from ref 47, while the last four are from ref 67. Results for fephen2ncs2 are taken from ref 48. The majority of the complexes are ground state singlet Fe(II)−N6 (most with aromatic nitrogens) crossing over to a high-spin quintet thereby changing the number of unpaired electrons by four.

contrast, DBLOC provides systematically appropriate corrections, in sign and magnitude, across the entire suite of test cases. The reason for the failure of B3LYP* for these test cases is straightforward. When applied to the "standard" set of spin-crossover complexes, the B3LYP errors are generally in the 5−10 kcal/mol range and the high-spin state is overstabilized due to reduction of $\Delta_o$ as discussed previously. The presence of a $e_g \rightarrow t_{2g}$ transition between the ground and relevant excited spin states, and the numerical value of the error in the energy gap, are a consequence of the metals in these compounds being primarily Fe or Co and the ligands virtually all being aromatic nitrogens, as noted above. These are the cases for which B3LYP* was optimized, and subsequent "tests" of the method employed cases that differ very little in the aforementioned key respects. The test cases we have presented, in which B3LYP* performs poorly, in contrast involve different metals and ligands, and in many cases involve movement of the electron exclusively in the $t_{2g}$ or $e_g$ manifold. The overfitting of B3LYP* to a narrowly defined training set then becomes apparent, and its unsuitability for general use manifest.

**3.8. Applications to Small-Gap Spin-Crossover Complexes.** Molecular orbital diagrams of a number of small-gap spin-crossover octahedral complexes from the literature are shown in Table 10 (see Supporting Information Table 14 for models of complexes). The first 9 complexes are from ref 47, the fephen2ncs2 complex is from ref 48. and the last 4 complexes are from ref 67. Some of the complexes from the last reference have not been included here because the DFT results reported show heavy dependence on both functional and basis set and there are not enough calculation details provided to resolve this. Most of the complexes are low-spin singlet ground state Fe(II)−N6 (and most of the N atoms are aromatic) involving a spin-crossover to the high-spin quintet, where there are 4 more unpaired electrons. All of these complexes have either one or two $t_{2g} \rightarrow e_g$ nonradiative transitions.

Table 11 shows B3LYP and experimental spin-crossover energetics for the small-gap complexes. The B3LYP results reported here are simply those values reported in the literature and vibrational relaxation effects do not need to be taken into account given the experimental protocols used in these cases. Details of the B3LYP calculations and the experimental results are described in the literature.[47,48,67] In every case conventional B3LYP incorrectly reverses the ordering of the spin states with respect to experiment, predicting the high-spin state to be the ground state by as much as 8 kcal/mol. This is in agreement with the observation that B3LYP tends to overbind high-spin states and hence these are the ground states for complexes with near zero gaps. Experimental gaps are all within about 5 kcal/mol. On average the conventional B3LYP spin-crossover energies have errors

with respect to experiment of 8.39 kcal/mol with a standard deviation of 3.21 kcal/mol and a maximum error of 12.43 kcal/mol.

Using the literature B3LYP and experimental results shown in Table 11 for the spin-crossover energy shows the DBLOC models necessity of the exmss parameter of 2.85 kcal/mol for aromatic nitrogen in correcting the small-gap spin-crossover data. DBLOC corrects the B3LYP ordering of the spin states in all but 2 important cases (fetacn2 and febptnncs2). The MUE for DBLOC is 3.51 kcal/mol with a standard deviation of 2.98 kcal/mol and a maximum of 10.16 kcal/mol. This MUE and maximum error might be substantially decreased if we were to recompute the B3LYP DFT results using our standard basis sets and convergence protocol, an analysis we plan to carry out in the near future. As it is, the reduction of MUE from 8.39 to 3.51 kcal/mol, with no further adjustment of parameters and no recomputation of DFT with a single consistent basis set, represents a very substantial improvement.

## 4. Conclusion

A simplified five-parameter empirical correction scheme based on ligand field theory is proposed to give more accurate energies for B3LYP calculations on relative spin state energetics of transition metal complexes. By isolating errors in radiative spin-forbidden transition energies over a diverse database of octahedral transition metal complexes some general rules about the behavior of B3LYP in comparison to experiment can be made. Errors in the treatment of nondynamical correlation of occupied d-orbitals results in B3LYP's large underestimation, 10.05 kcal/mol, of the energy required to pair two $t_{2g}$ or $e_g$ electrons. Additionally, errors in $\Delta_o$ were found to be proportional to the strength of the metal−ligand interaction, that is, smaller, 1.88 kcal/mol, intermediate, 2.85 kcal/mol, and larger, 5.21 kcal/mol, corrections to the left, middle, and right of the spectrochemical series. B3LYP also slightly overestimates parallel spin−spin interactions in metal complexes. The DBLOC model brings the average error in the spin-splitting from 10.14 to 1.98 kcal/mol. Comparison of B3LYP, B3LYP*, and DBLOC clearly shows the advantages of using DBLOC over B3LYP and B3LYP* (15% exact nonlocal exchange) for which the spin-forbidden transition errors are often worse than B3LYP by about 5 kcal/mol. Applying DBLOC to the seven complexes from Swart et al.[66] shows that it is in agreement with experiment. Applying DBLOC to spin-crossover compounds with near zero gaps reduces the MUE of conventional B3LYP from 8.39 to 3.51 kcal/mol.

**Acknowledgment.** This work has been supported by the Department of Energy program through solar photochemistry (DE-FG02-90ER-14162) to R.A.F.

**Supporting Information Available:** Tables of B3LYP equilibrium geometries for the various DBLOC metal complexes in their different spin states, along with discussion of these results, experimental results for $t_{2g} \rightarrow e_g$ and $e_g \rightarrow t_{2g}$ spin-forbidden transitions, tables and discussion describing the application of the DBLOC model to complexes from Swart et al.,[66] additional tables including models of complexes, DBLOC database descriptors, and supplementary vibrational relaxation data, and Cartesian coordinates of all complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Rinaldo, D.; Philipp, D.; Lippard, S.; Friesner, R. *J. Am. Chem. Soc.* **2007**, *129*, 3135.

(2) Concepcion, J.; Jurss, J.; Brennaman, M.; Hoertz, P.; Patrocinio, A.; Iha, N.; Templeton, J.; Meyer, T. *Acc. Chem. Res.* **2009**, *42*, 1954.

(3) Gutlich, P.; Hauser, A.; Spiering, H. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 2024.

(4) Harvey, J. N.; Aschi, M. *Faraday Discuss.* **2003**, *124*, 129.

(5) Carreón-Macedo, J.-L.; Harvey, J. N. *Phys. Chem. Chem. Phys.* **2006**, *8*, 93.

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

(7) Ballhausen, C. J. *Introduction to Ligand Field Theory*; McGraw-Hill Book Company, Inc.: New York, 1962.

(8) Douglas, B.; McDaniel, D.; Alexander, J. *Concepts and Models of Inorganic Chemistry*; John Wiley & Sons Inc.: New York, 1994.

(9) Steinfeld, J. I. *Molecules and Radiation: An Introduction to Modern Molecular Spectroscopy*; The MIT Press, Inc.: Cambridge, MA, 1985.

(10) Harris, D. C.; Bertolucci, M. D. *Symmetry and Spectroscopy: An Introduction to Vibrational and Electronic Spectroscopy*; Oxford University Press, Inc.: New York, 1978.

(11) Shimura, Y. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 693.

(12) Friesner, R.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.

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

(14) Cramer, C.; Truhlar, D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757.

(15) Rinaldo, D.; Tian, L.; Harvey, J.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 164108.

(16) Nolet, M.-C.; Beaulac, R.; Boulanger, A.-M.; Reber, C. *Struct. Bonding* **2004**, *107*, 145.

(17) Allen, G. C.; El-Sharkawy, G. A. M. *Inorg. Nucl. Chem. Lett.* **1970**, *6*, 493.

(18) Vanquickenborne, L. G.; Coussens, B.; Postelmans, D.; Ceulemans, A.; Pierloot, K. *Inorg. Chem.* **1991**, *30*, 2978.

(19) Friesen, D. A.; Nashiem, R. E.; Waltz, W. L. *Inorg. Chem.* **2007**, *46*, 7982.

(20) Choi, J.-H.; Choi, S. Y.; Hong, Y. P.; Ko, S.-O.; Ryoo, K. S.; Lee, S. H.; Park, Y. C. *Spectrochim. Acta, Part A* **2008**, *70*, 619.

(21) Berben, L. A.; Long, J. R. *Inorg. Chem.* **2005**, *44*, 8459.

(22) Tsukahara, Y.; Kamatani, T.; Lino, A.; Suzuki, T.; Kaizaki, S. *Inorg. Chem.* **2002**, *41*, 4363.

(23) Stewart J. J. P. Mopac. http://openmopac.net/manual/Transition_metal_complexes.html, 2010.

(24) Jørgensen, C. K. *Inorg. Chim. Acta* **1969**, *3*, 313.

(25) Arulsamy, N.; Hodgson, D. J. *Inorg. Chem.* **1994**, *33*, 4531.

(26) Manson, N. B.; Shah, G. A.; Howes, B.; Flint, C. D. *Mol. Phys.* **1977**, *34*, 1157.

(27) Childers, M. L.; Su, F.; Przyborowska, A. M.; Bishwokarma, B.; Park, G.; Brechbiel, M. W.; Torti, S. V.; Torti, F. M.; Broker, G.; Alexander, J. S.; Rogers, R. D.; Ruhlandt-Senge, K.; Planalp, R. P. *Eur. J. Inorg. Chem.* **2005**, *19*, 3971.

(28) Baho, N.; Zargarian, D. *Inorg. Chem.* **2007**, *46*, 299.

(29) Nolet, M.-C.; Michaud, A.; Bain, C.; Zargarian, D.; Reber, C. *Photochem. Photobiol.* **2006**, *82*, 57.

(30) González, E.; Rodrigue-Witchel, A.; Reber, C. *Coord. Chem. Rev.* **2007**, *251*, 351.

(31) Buñuel, M. A.; García, J.; Proietti, M. G.; Solera, J. A.; Cases, R. *J. Chem. Phys.* **1999**, *110*, 3566.

(32) Bussiére, G.; Reber, C.; Neuhauser, D.; Walter, D. A.; Zink, J. I. *J. Phys. Chem. A* **2003**, *107*, 1258.

(33) Reimann, C. W. *J. Phys. Chem.* **1970**, *74*, 561.

(34) Jørgensen, C. K. *Acta Chem. Scand.* **1955**, *9*, 1362.

(35) Ewald, A. H.; Martin, R. L.; Ross, I. G.; White, A. H. *Proc. Royal Soc. A* **1964**, *280*, 235.

(36) Renovitch, G. A.; Baker, W. A. *J. Am. Chem. Soc.* **1968**, *90*, 3585.

(37) Naiman, C. S. *J. Chem. Phys.* **1961**, *35*, 323.

(38) Gray, H. B.; Beach, N. A. *J. Am. Chem. Soc.* **1963**, *85*, 2922.

(39) Karpishin, T. B.; Gebhard, M. S.; Solomon, E. I.; Raymond, K. N. *J. Am. Chem. Soc.* **1991**, *113*, 2977.

(40) Maiti, D.; Paul, H.; Chanda, N.; Chakraborty, S.; Mondal, B.; Puranik, V. G.; Lahiri, G. K. *Polyhedron* **2004**, *23*, 831.

(41) Sharrad, C. A.; Lüthi, S. R.; Gahan, L. R. *Dalt. Trans.* **2003**, *19*, 3693.

(42) Goodall, D. M.; Hollis, D. B.; White, M. S. *J. Phys. Chem.* **1987**, *91*, 4255.

(43) Bruce, J. I.; Gahan, L. R.; Hambley, T. W.; Stranger, R. *Inorg. Chem.* **1993**, *32*, 5997.

(44) Donlevy, T. M.; Gahan, L. R.; Hambley, T. W.; McMahon, K. L.; Stranger, R. *Aust. J. Chem.* **1993**, *46*, 1799.

(45) J. Ferguson, D. L. W.; Knox, K. *J. Chem. Phys.* **1963**, *39*, 881.

(46) Deeth, R. J.; Anastasi, A. E.; Wilcockson, M. J. *J. Am. Chem. Soc.* **2010**, *132*, 6876.

(47) Jensen, K. P.; Cirera, J. *J. Phys. Chem. A* **2009**, *113*, 10033.

(48) Reiher, M. *Inorg. Chem.* **2002**, *41*, 6928.

(49) Zhao, Y.; Truhlar, D. G *J. Chem. Phys.* **2006**, *125*, 194101.

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

(51) Salomon, O.; Reiher, M.; Artur Hess, B. *J. Chem. Phys.* **2002**, *117*, 4729.

(52) Friesner, R. A.; Knoll, E. H.; Cao, Y. *J. Chem. Phys.* **2006**, *125*, 124107.

(53) Knoll, E. H.; Friesner, R. A. *J. Phys. Chem. B* **2006**, *110*, 18787.

(54) Goldfeld, D.; Bochevarov, A.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 214105.

(55) Hall, M.; Goldfeld, D.; Bochevarov, A.; Friesner, R. A. *J. Chem. Theory Comput.* **2009**, *5*, 2996.

(56) Allen, F. H. *Acta Crystallogr.* **2002**, *B58*, 380.

(57) *Jaguar*, version 7.5; Schrödinger, Inc.: New York, NY, 2009.

(58) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.

(59) Vacek, G.; Perry, J. K.; Langlois, J.-M. *Chem. Phys. Lett.* **1999**, *310*, 189.

(60) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737.

(61) Berger, R.; Fischer, C.; Klessinger, M. *J. Phys. Chem. A* **1998**, *102*, 7157.

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

(63) Liao, M.-S.; Huang, M.-J.; Watts, J. D. *J. Phys. Chem. A* **2010**, *114*, 9554.

(64) Engauge Digitizer, http://digitizer.sourceforge.net, 2010.

(65) Berning, A.; Schweizer, M.; Werner, H.-J.; Knowles, P. J.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1823.

(66) Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479.

(67) Paulsen, H.; Duelund, L.; Winkler, H.; Tolflund, H.; Trautwein, A. X. *Inorg. Chem.* **2001**, *40*, 2201.

CT100359X

JCTC Journal of Chemical Theory and Computation

# Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated *ab initio* Methods?

Frank Neese* and Edward F. Valeev

*Lehrstuhl für Theoretische Chemie, Universität Bonn, Wegelerstr. 12, D-53115 Bonn, Germany, and Department of Chemistry, Virginia Tech, Blacksburg, Virginia 24061-0001, United States*

Received July 17, 2010

**Abstract:** The performance of several families of basis sets for correlated wave function calculations on molecules is studied. The widely used correlation-consistent basis set family cc-pVXZ ($n$ = D, T, Q, 5) is compared to a systematic series of atomic natural orbital basis sets (ano-pVXZ). These basis sets are built from the cc-pV6Z primitives in atomic multireference average coupled pair functional (MR-ACPF) calculations. Segmented basis sets optimized for self-consistent field calculations (def2-SVP, def2-TZVPP, and def2-QZVPP as well as "pc-*n*", $n$ = 1, 2, 3) were also tested. Reference Hartree−Fock energies are determined with the uncontracted aug-cc-pV6Z basis set for a set of 21 small molecules built from H, B, C, N, O, and F. Reference coupled cluster CCSD(T) correlation energies were determined from extrapolation at the cc-pV5Z/cc-pV6Z level. It is found that the ano-pVXZ basis sets outperform the other basis sets. The error in the SCF energies compared to cc-pVXZ basis sets is reduced by about a factor of 3 at each cardinal number. In addition, the ano-pVXZ consistently recovers more correlation energy than their competitors at each cardinal number. The ability of the four families of basis sets to extrapolate SCF and correlation energies to the basis set limit has been investigated. A conclusion by Truhlar is confirmed that the optimum exponent for correlation energy extrapolations at the DZ/TZ level is ∼2.4. All TZ/QZ basis set pairs lead to an optimum exponent close to the expected value of 3. The SCF energy extrapolation proposed by Petersson and co-workers is found to be effective. At the DZ/TZ level, errors in *total* energies of less than 2 mEh are found for the test set, while at the TZ/QZ level one obtains the total energies within ∼0.3 mEh of the basis set limit. For extrapolation, the "cc" and "ano" bases are found to be similarly successful. Extrapolation results were compared to explicitly correlated calculations with dedicated basis sets (cc-pVXZ-F12) as well as the ano-pVXZ bases. It is found that the ano-pVXZ+ basis sets perform as well as the cc-pVXZ-F12 family (both are of comparable size); additional improvement should be possible by reoptimizing the ANO basis sets for explicitly correlated calculations. The error of the extrapolated energies is about 2−3 times smaller than what was found in the explicitly correlated calculations. However, the error in the explicitly correlated calculations is more systematic, and hence the same conclusion may not hold for the computation of energy differences.

## 1. Introduction

It is well-known that the one-particle basis set is an important ingredient of wave-function-based *ab initio* calculations on molecules. While the convergence of the Hartree−Fock energy to the basis set limit is relatively rapid as the one-particle basis set is approaching completeness, the convergence of the correlation energy is known to be slow.[1] In order to obtain accurate results for energetic quantities, very large basis sets of at least polarized quadruple-ζ quality or even larger are required. Such basis sets quickly become

* Corresponding author e-mail: neese@thch.uni-bonn.de.

10.1021/ct100396y 
Published on Web 12/03/2010

unmanageably large even for relatively small molecules (say, beyond five non-hydrogen atoms).

In recent years, two techniques have emerged that aim at obtaining accurate results for correlated *ab initio* energies with smaller basis sets. The first family of methods corrects the origin of the slow basis set convergence of the correlation energy—the lack of derivative discontinuity (cusp) in standard wave functions when electrons meet each other. The *explicitly correlated* methods accomplish this by introducing the interelectronic distances $r_{ij}$ explicitly into the wave function *ansatz*; this can be viewed alternatively as an inclusion of $r_{ij}$-dependent many-electron basis functions. The most practical of these methods were pioneered by Kutzelnigg;[2] the *R12 methods* developed by him and many others allow for approaching the basis set limit with much smaller basis sets. Great progress has been made in recent years in the formulation and implementation of explicitly correlated wave function methods. An authoritative review of these approaches is provided in a recent article by Helgaker et al.[3]

The second family of methods relies on extrapolation of the correlation energy based on a series of calculations with successively refined basis sets.[4,5] There are two key requirements for this approach: (a) the existence of a systematically converging series of basis sets and (b) a suitable formula for correlation energy extrapolation.

Dunning and co-workers made a major contribution by developing a series of successively larger one-particle basis sets, cc-pVXZ with $n = 2$ (= D), 3 (= T), 4 (= Q), 5, 6, etc. being the "cardinal number"; cc stands for "correlation consistent" and p for "polarized".[6−15] Using these basis sets, the calculated correlation energies (within a given correlation method) converge smoothly and systematically to the basis set limit. The cc-basis sets are built upon a core of Hartree−Fock orbitals and are systematically supplemented with additional primitive Gaussian functions that describe polarization and valence shell correlation. For $n = 2$, 3, 4, etc., the polarization sets for main group elements consist of 1d, 2d1f, 3d2f1g, etc. functions. For the hydrogen and helium atoms, the equivalent sets consist of 1p, 2p1d, 3p2d1f, etc. functions. Further modifications include additional diffuse functions for weak interactions and calculations on anions (aug-cc-pVXZ)[10] and functions that describe core-correlation (cc-pCVXZ[9,16]) or core−valence correlation (cc-pwCVXZ[6,9]). Recently, Peterson and co-workers have developed correlation consistent basis sets for first through third row transition metals.[17−20] Recontractions of the basis sets to comply with scalar relativistic Douglas−Kroll−Hess (DKH) calculations are also available.

A second way to construct systematic series of one-particle basis sets was pioneered by Almlöf and Taylor and is based on atomic natural orbitals (ANOs).[18−22] ANO basis sets are generally contracted in the sense that all Gaussian primitives of a given angular momentum contribute to all basis functions. ANOs are among the best possible choices for atomic correlation calculations. It is, however, not a priori clear whether ANOs are flexible enough to properly describe low symmetry molecular environments and the changes of the atomic orbitals upon bond formation and charge transfer. Many excellent results have, however, been obtained with the use of ANOs, and they are the main workhorse of the MOLCAS quantum chemistry program.[21−23] The *general contraction* scheme on ANOs poses new challenges on the integral evaluation program: the two-electron integral evaluation time naively depends on the fourth power of the contraction depth, although the scaling is reduced to quadratic by using the robust density fitting (aka resolution of the identity, RI) approach.[24] Only a few modern quantum chemistry programs can efficiently handle generally contracted basis sets.

In SCF calculations, the computing time is strongly dominated by the integral evaluation time, and hence one seeks to develop basis sets with the smallest possible number of primitives. This has been the route followed by Ahlrichs and co-workers in the development of the def- series of basis sets that is available in double-ζ (def2-SVP), triple-ζ (def2-TZVPP), and quadruple-ζ (def2-QZVPP) variants.[28−32] For these basis sets, exponents and contraction coefficients are optimized in atomic calculations. The def2 bases usually show excellent performance in Hartree−Fock and DFT studies. The behavior of these basis sets in MP2 calculations has been studied by their developers,[25,26] but the behavior in highly correlated calculations has probably not been fully assessed to date.

A similar target has been followed with the development of the pc-$n$ ($n = 0$, 1, 2, 3, 4) basis sets by Jensen and co-workers.[27−33] The basis sets of double- through quintuple-ζ quality were designed to systematically converge to the SCF basis set limit. These basis sets are partially optimized in molecular calculations.

The motivation for the present study was the efficient application of *accurate* correlation methods for large molecules. Since the overall cost is dominated by the calculation of the correlation energy for which the integral evaluation time is less of an issue, the most important aspect is the number of basis functions with which a given accuracy can be obtained. Clearly, the error relative to the basis set limit contains contributions from the SCF error and the correlation energy error. Hence, one desires basis sets that, for a given size, perform well in both respects. In this context, we became somewhat unsatisfied with the correlation-consistent basis sets that provide excellent correlation energies but also show errors in the Hartree−Fock energies that are significantly larger than what is obtained with other basis sets of the same size. We were curious of whether one could improve on this behavior while maintaining good correlation energies. This is also relevant in the context of explicitly correlated calculations where one aims at accurate correlation energies with small basis sets. The present work represents an attempt to systematically evaluate the performance of series of basis sets for which at least double-, triple-, and quadruple-ζ variants are available. Particularly important are the double- and triple-ζ members within each series since calculations with larger basis sets are hardly feasible in "real life" chemical applications. Since errors can be largely reduced through extrapolation and R12 techniques, we have also evaluated the performance of the various basis sets with respect to SCF and correlation energy extrapolations using these two approaches.

## 2. Basis Set Construction

A series of ANO basis sets was constructed in the course of this study. It is obvious that no basis set can be more accurate than the underlying set of primitives. Hence, we have used the primitives of the cc-pV6Z basis set[34] as a starting point. In order to keep the calculations manageable, a two-step procedure was pursued.

The ANO basis sets were obtained from calculations on the neutral atoms. To this end, the cc-pV6Z basis set was fully decontracted. Atomic CASSCF calculations were carried out for the atomic ground terms. The atomic calculations were tightly converged on the atomic ground terms while carefully averaging over the spatially degenerate member of the terms. For example, for the $^3$P state of the carbon atom, all three triplet roots corresponding to the three components of the P-state were determined. Multireference averaged coupled pair functional (MR-ACPF[15]) calculations were performed on top of the CASSCF ground states. The MR-ACPF method has the advantage over MR configuration interaction (MR-CI) of being (nearly) size consistent and leading to a stationary solution with a well-defined density.[35] For the MR-ACPF calculations, all configurations in the CAS space were kept as references, even those of the "wrong" symmetry and insufficient number of open-shell orbitals. The first-order interacting space is left uncontracted, and no selection or any other approximation was made. The densities of all three components of the P states were averaged in order to obtain a spherically symmetric density that was subsequently diagonalized in order to obtain atomic natural orbitals. All valence electrons were correlated. Thus, the cores of our ANO basis sets are orbitals optimized at the CASSCF level.

Three families of basis sets were constructed in order to investigate the transition from cc-basis sets to ANO basis sets:

(1) acc-pVXZ: The original polarization functions of the cc-pVXZ basis were kept, and only the s and p functions were replaced by their ANO counterparts. The number of primitives is 16 for the s functions and 12 for the p functions (B−Ne).

(2) rcc-pVXZ: the same s and p contractions as in cc-pVXZ were used, but the polarization functions were replaced by ANO contractions. The same numbers of contracted polarization functions were kept as in the cc-pVXZ family, e.g., 1d for $n = 2$, 2d1f for $n = 3$, and 3d2f1g for $n = 4$. The number of primitives is five for the d contractions, four for the f contractions, and three for the g contractions.

(3) ano-pVXZ: All orbitals were left at their complete ANO contractions as described above. For hydrogen, the appropriate number of ANOs is taken from the original NASA-AMES ANO set.[35,36]

The hydrogen basis for the ano-pVXZ and acc-pVXZ basis sets was taken from the original NASA-AMES-ANO basis,[37] while in keeping with the philosophy of the basis set construction, the *s and p* part for the rcc bases came from the original cc-bases and the polarization functions from the NASA-AMES set.

Thus, by construction the four families of basis sets cc-pVXZ, acc-pVXZ, rcc-pVXZ, and ano-pVXZ all have the same number of basis functions for a given $n$ but differ in their contraction depth.

For comparison, the segmented basis sets def2-SVP, def2-TZVPP, and def2-QZVPP developed by Ahlrichs and co-workers as well as the polarization consistent basis sets of Jensen (pc-$n$, $n = 1, 2, 3$) were included in the study. Furthermore, alternative ANO sets due to Roos and co-workers[23,36] as well as the original, pioneering quadruple-ζ type ANO basis by Almlöf and Taylor[24] were included in the study. The number of basis functions is comparable or even identical to the other basis sets included in the study, and hence a comparison is appropriate.

## 3. Calculations

Twenty-seven basis sets have been evaluated by performing molecular calculations. In order to judge the performance of the basis sets, the Hartree−Fock energy as well as the correlation energy at the CCSD(T) level will be employed. The SCF energies calculated with the very large uncontracted aug-cc-pV6Z basis set serve as a near HF limit reference. We expect the deviation of the reference values from the true Hartree−Fock limit to be below 0.1 mEh. Reference values for the correlation energies were obtained by using the standard two-point extrapolation scheme involving the cc-pV5Z and cc-pV6Z basis sets, as will be detailed below.

As test systems, the hydrides $H_2$, $BH_3$, $CH_4$, $NH_3$, $H_2O$, and FH were chosen together with all possible diatomics built from B, C, N, O, and F. Open-shell species were treated in the spin-unrestricted formalism. A reviewer pointed out that we have inadvertedly used an excited state of BN. For the purposes of this paper, this choice is immaterial. The geometries were taken from geometry optimizations at the B3LYP/def2-TZVP level. Geometric parameters are given in the Computational Details section.

## 4. Comparison of Basis Sets

The reference values used in the present study are collected in Table 1. The results of the test calculations are summarized in terms of statistical measures in Table 2. Individual results for all molecules and basis sets are given in the Supporting Information.

**4.1. SCF Energies.** The performance of the various double-ζ basis sets varies fairly dramatically. The largest absolute errors are obtained with the def2-SVP basis set that has been designed with the smallest possible number of primitive Gaussians in mind. Hence, calculations with def2-SVP are very efficient, and many successful molecular studies have been performed with this basis set. Nevertheless, it is the one with the largest deviations from the Hartree−Fock limit in the present study. Perhaps surprisingly, the second largest errors are obtained from the pc-2 basis set that has also been designed for SCF (Hartree−Fock and DFT calculations). Third in line is the cc-pVDZ basis set that shows a mean unsigned deviation from the basis set limit of about 40 mEh in the present test set.

***Table 1.*** Reference SCF and Correlation Energies for the Present Study

| molecule | SCF-reference | EC(CCSD(T)) cc-pV5Z | EC(CCSD(T)) cc-pV6Z | E(CCSD(T)) cc-pV(5/6)Z | Etot(CCSD(T)) |
|---|---|---|---|---|---|
| $H_2$ ($^1\Sigma$) | −1.133583 | −0.04065 | −0.04076 | −0.04090 | −1.17449 |
| $BH_3$ ($^1A$) | −26.402596 | −0.14446 | −0.14505 | −0.14586 | −26.54846 |
| $CH_4$ ($^1A$) | −40.216965 | −0.23767 | −0.23884 | −0.24044 | −40.45741 |
| $NH_3$ ($^1A$) | −56.224824 | −0.27479 | −0.27660 | −0.27909 | −56.50392 |
| $H_2O$ ($^1A$) | −76.066958 | −0.30233 | −0.30499 | −0.30866 | −76.37561 |
| HF ($^1A$) | −100.070300 | −0.31511 | −0.31874 | −0.32372 | −100.39402 |
| $B_2$ ($^3\Sigma$) | −49.086197 | −0.17822 | −0.17885 | −0.17972 | −49.26592 |
| BC ($^4\Sigma$) | −62.337076 | −0.19774 | −0.19887 | −0.19970 | −62.53678 |
| BN ($^3\Sigma$) | −79.021852 | −0.27087 | −0.27280 | −0.27547 | −79.29732 |
| BO($^2\Sigma$) | −99.566250 | −0.34236 | −0.34539 | −0.34955 | −99.91580 |
| BF ($^1\Sigma$) | −124.168631 | −0.38238 | −0.38632 | −0.39172 | −124.56035 |
| $C_2$ ($^1\Sigma$) | −75.406555 | −0.39968 | −0.401452347 | −0.40275 | −75.80930 |
| CN ($^2\Sigma$) | −92.234733 | −0.35337 | −0.35581 | −0.35916 | −92.59389 |
| CO ($^1\Sigma$) | −112.791245 | −0.40711 | −0.41052 | −0.41521 | −113.20646 |
| CF ($^2\Pi$) | −137.238721 | −0.41972 | −0.42407 | −0.43004 | −137.66876 |
| $N_2$ ($^1\Sigma$) | −108.994314 | −0.42002 | −0.42329 | −0.42778 | −109.42209 |
| NO ($^2\Pi$) | −129.309819 | −0.45716 | −0.46120 | −0.46675 | −129.77657 |
| NF ($^3\Sigma$) | −153.852696 | −0.45705 | −0.46194 | −0.46867 | −154.32136 |
| $O_2$ ($^3\Sigma$) | −149.691600 | −0.49790 | −0.50271 | −0.50931 | −150.20090 |
| OF ($^2\Pi$) | −174.210823 | −0.53157 | −0.53748 | −0.54560 | −174.75642 |
| $F_2$ ($^1\Sigma$) | −198.775205 | −0.60622 | −0.61330 | −0.62302 | −199.39823 |

It is interesting how much better one can get with an ANO basis set. In fact, the error with the ano-pVDZ basis is more than a factor of 3 smaller than what is achieved with cc-pVDZ and an order of magnitude smaller than the error obtained with the def2-SVP basis. This accuracy is obtained despite the fact that both basis sets contain the same number of contracted basis functions. The comparison of the ano-pVDZ, acc-pVDZ, rcc-pVDZ, and cc-pVDZ basis sets reveals that the improvement in the SCF energy is mainly due to the highly contracted ANO s and p parts of the basis set and to a lesser extent to the contracted single polarization function.

Already at the triple-$\zeta$ level with 2d1d and 2p1d polarization sets, a more uniform performance of the various basis sets is obtained. The basis set that leads to the highest energies is the cc-pVTZ one, closely followed by pc-2. The performance of the def2-TZVPP basis set is excellent, and it only gives a mean unsigned error of about 5 mEh. However, once more, the improvement obtained with the ano-pVTZ basis set is highly significant. The error is again more than a factor of 3 smaller than that obtained with the cc-pVTZ basis and only amounts to about 2.5 mEh on average. This is almost as good as the result obtained with the much larger cc-pVQZ basis.

For the quadruple-$\zeta$ bases, all basis sets provide results within 2.5 mEh from the reference values. Once more, the by far largest error is obtained with cc-pVQZ, followed by acc-pVQZ, which has the same s and p functions. The best results are obtained with def2-QZVPP, which comes within 0.6 mEh of the reference values. In this respect, it is marginally better than the ano-pVQZ basis set that shows an error that is 0.2 millihartree larger. However, def2-QZVPP is also slightly larger in terms of basis functions than cc-pVQZ or ano-pVQZ. The same is true for pc-3, which contains an extra set of s and p functions for B−Ne but is still slightly less accurate than ano-pVQZ.

It is instructive to observe that very significant improvements in the results can be obtained by merely adding an additional set of s and p functions to the basis set without extending the polarization set. This leads to sano-pVDZ+ (s stands for small in this case). For the sano-pVDZ+ basis set, the mean unsigned error in the SCF energy is only 7.0 mEh, which is almost half of the error obtained with ano-pVDZ and already better than what is obtained with the much larger cc-pVTZ or pc-2 bases. In fact, for the sano-pVTZ+ basis set where the four additional basis functions are even less problematic, the mean unsigned error drops to 1.4 mEh, which is about half of what one gets from the much larger cc-pVQZ basis. Even for the sano-pVQZ+ basis, an improvement is obtained, and a mean unsigned error of only 0.5 mEh results. Since sano-pVQZ+ is about the same size as def2-QZVPP, this shows that ANO basis sets are even competitive in terms of accuracy with dedicated SCF optimized bases, even though they have never been designed for this purpose.

By adding the next d function to the polarization set and an extra s function, one obtains ano-pVDZ+ (5s3p2d for B−Ne), which improves the results further, and for ano-pVTZ+, the average unsigned error even drops below 1 mEh. In terms of ANOs, one can also think about the ano-pVXZ+ basis sets as being identical to the ano-pV($n$+1)Z basis set with the highest angular momentum polarization function deleted and an extra s function added (the latter does not add significantly to the computational cost and mainly improves the SCF energies). Hence, these basis sets are significantly larger than the ano-pVXZ bases.

In a comparison of our results to the those of the ANO bases of Roos and co-workers and Almlöf and Taylor, we observe that the results are fairly similar, with slight advantages for the ano-pVXZ basis sets constructed in this work.

In summary, from the results collected in Table 2, one concludes that ANO basis sets yield excellent SCF energies. In fact, the results obtained with ano-pVXZ are almost as good as what one obtains from cc-pV($n$+1)Z and also as good or even better that what one obtains with basis sets that are specifically optimized for SCF calculations. This behavior was not anticipated and is important in the context of explicitly correlated calculations, as will be discussed below.

***Table 2.*** Errors (in mEh) in SCF and Correlation Energies Relative to the Uncontracted cc-pV6Z and Complete Basis Extrapolated Reference Values Respectively[a]

| basis set | MUE(SCF) | MUE(EC) | %$\Delta E_C$ | RMS(SCF) | RMS(EC) | MAX(SCF) | MAX(EC) | MUE($E_{tot}$) |
|---|---|---|---|---|---|---|---|---|
| cc-pVDZ | 39.9 | 98.1 | 26.1 | 45.7 | 110.8 | 88.2 | 213.3 | 138.0 |
| cc-pVTZ | 9.8 | 35.5 | 9.2 | 11.1 | 40.8 | 21.3 | 81.0 | 45.3 |
| cc-pVQZ | 2.2 | 14.6 | 3.7 | 2.5 | 17.0 | 5.0 | 34.4 | 16.7 |
| pc-1 | 74.4 | 100.9 | 27.2 | 84.2 | 113.7 | 158.8 | 222.6 | 175.2 |
| pc-2 | 8.3 | 43.7 | 11.4 | 9.1 | 50.1 | 15.4 | 100.3 | 52.0 |
| pc-3[b] | 0.9 | **(13.3)** | **(3.4)** | 0.9 | 15.5 | **(1.3)** | 31.5 | 14.2 |
| def2-SVP | 128.2 | 99.2 | 26.4 | 144.0 | 112.1 | 266.8 | 216.2 | 227.4 |
| def2-TZVPP | 4.9 | 36.9 | 9.6 | 5.4 | 42.4 | 9.0 | 84.1 | 41.8 |
| def2-QZVPP[c] | **(0.6)** | 15.6 | 4.0 | **(0.7)** | 18.2 | **(1.2)** | 36.8 | 16.2 |
| ano-pVDZ | **13.1** | **82.6** | **21.9** | **15.0** | **92.9** | **29.0** | **173.0** | **95.7** |
| ano-pVTZ | **2.6** | **30.4** | **7.9** | **2.8** | **34.9** | **4.5** | **65.9** | **33.0** |
| ano-pVQZ | **0.8** | **12.9** | **3.3** | **0.9** | **15.0** | **1.7** | **28.4** | **13.7** |
| acc-pVDZ | 16.6 | 92.5 | 24.5 | 18.3 | 104.7 | 30.2 | 202.2 | 109.2 |
| acc-pVTZ | 3.7 | 35.6 | 9.3 | 4.1 | 40.5 | 7.0 | 78.5 | 39.3 |
| acc-pVQZ | 1.1 | 15.2 | 3.9 | 1.2 | 17.6 | 2.3 | 35.6 | 16.2 |
| rcc-pVDZ | 36.3 | 89.1 | 23.7 | 42.0 | 100.5 | 83.1 | 192.9 | 125.5 |
| rcc-pVTZ | 9.0 | 32.1 | 8.3 | 10.4 | 36.8 | 20.3 | 72.4 | 41.1 |
| rcc-pVQZ | 2.0 | 13.4 | 3.4 | 2.4 | 15.5 | 4.8 | 31.3 | 15.4 |
| larger default basis sets | | | | | | | | |
| aug-cc-pVDZ | 32.4 | 84.1 | 22.4 | 37.3 | 94.4 | 75.0 | 177.1 | 116.5 |
| aug-cc-pVTZ | 8.2 | 29.8 | 7.8 | 9.4 | 34.1 | 18.6 | 66.3 | 38.0 |
| aug-cc-pVQZ | 1.7 | 12.2 | 3.1 | 2.0 | 14.2 | 4.4 | 28.3 | 14.0 |
| sano-pVDZ+ | 7.0 | 67.3 | 17.6 | 7.9 | 76.4 | 13.1 | 143.3 | 74.3 |
| sano-pVTZ+ | 1.4 | 12.4 | 6.9 | 1.5 | 30.9 | 2.5 | 58.5 | 28.2 |
| sano-pVQZ+ | 0.5 | 11.9 | 3.1 | 0.5 | 13.8 | 0.9 | 26.5 | 12.3 |
| ano-pVDZ+ | 5.0 | 53.4 | 13.7 | 5.7 | 61.6 | 11.0 | 115.4 | 58.4 |
| ano-pVTZ+ | 0.9 | 20.9 | 5.4 | 1.0 | 24.1 | 2.0 | 47.6 | 21.8 |
| ano-pVQZ+ | 0.2 | 8.7 | 2.3 | 0.3 | 9.9 | 0.5 | 19.5 | 8.9 |
| Roos-ANO-DZP[d] | 6.7 | 65.8 | 17.6 | 7.7 | 73.4 | 15.7 | 133.5 | 72.5 |
| Roos-ANO-TZP[d] | 1.4 | 23.3 | 6.1 | 1.6 | 26.6 | 2.7 | 51.0 | 24.6 |
| NASA-AMES-ANO[e] | 1.1 | 12.4 | 3.2 | 1.2 | 14.3 | 1.6 | 28.3 | 13.4 |
| explicitly correlated (R12) CCSD(T)[e] | | | | | | | | |
| ano-pVDZ | **13.1** | 22.1 | 6.5 | **15.0** | 25.4 | **29.0** | 51.7 | 23.1 |
| ano-pVTZ | **2.6** | 7.5 | 2.1 | **2.8** | 9.1 | **4.5** | 19.7 | 7.7 |
| ano-pVQZ | **0.8** | 2.8 | 0.8 | **0.9** | 3.4 | **1.7** | 7.6 | 2.8 |
| ano-pVDZ+ | 5.0 | 11.5 | 3.3 | 5.7 | 13.3 | 11.0 | 26.9 | 11.8 |
| ano-pVTZ+ | 0.9 | 3.6 | 1.0 | 1.0 | 4.4 | 2.0 | 9.5 | 3.6 |
| ano-pVQZ+ | 0.2 | 1.6 | 0.5 | 0.3 | 1.9 | 0.5 | 4.0 | 1.6 |
| cc-pVDZ | 39.9 | 25.1 | 7.6 | 45.7 | 28.3 | 88.2 | 55.3 | 27.8 |
| cc-pVTZ | 9.8 | 9.0 | 2.5 | 11.1 | 10.9 | 21.3 | 23.5 | 9.5 |
| cc-pVQZ | 2.2 | 3.7 | 1.0 | 2.5 | 4.6 | 5.0 | 10.2 | 3.9 |
| aug-cc-pVDZ | 32.4 | 16.0 | 4.8 | 37.3 | 17.7 | 75.0 | 32.7 | 17.2 |
| aug-cc-pVTZ | 8.2 | 6.1 | 1.7 | 9.4 | 7.3 | 18.6 | 15.7 | 6.2 |
| aug-cc-pVQZ | 1.7 | 2.7 | 0.7 | 2.0 | 3.4 | 4.4 | 7.6 | 2.7 |
| cc-pVDZ-F12 | 10.8 | 13.5 | 3.8 | 12.3 | 16.1 | 22.8 | 34.8 | 13.9 |
| cc-pVTZ-F12 | 2.0 | 5.0 | 1.4 | 2.3 | 5.8 | 4.7 | 11.9 | 5.1 |
| cc-pVQZ-F12 | 0.2 | 1.7 | 0.5 | 0.2 | 2.0 | 0.3 | 4.1 | 1.8 |

[a] The best result in each category of basis set (DZ, TZ, QZ) is printed in bold. If the best result is obtained with a slightly larger basis set, it is put in parentheses. MUE = mean unsigned error; RMS = root-mean square error; MAX = maximum error; SCF = Hartree-Fock energy; EC = CCSD(T) correlation energy. [b] The pc-3 basis set contains 6s5p4d2f1g contractions for B−Ne and is therefore the same size as sano-pVQZ+. [c] The def2-QZVPP basis set contains 7s4p3d2f1g contractions and is therefore of similar size to sano-pVQZ+. [d] The Roos-ANO basis sets contain the same number of functions as the ano-pVDZ+ and ano-pVTZ+ basis sets. NASA-AMES ANO is the same size as ano-pVQZ. [e] The CABS correction to the SCF energies have been included in the right-most column.

**4.2. Correlation Energies.** The main purpose of this study is to evaluate how the various basis sets perform in correlation energy calculations. As is evident from Table 2, the behavior of the basis sets is more uniform in this respect. With the DZ bases, one recovers 70−80% of the correlation energy; with TZ bases, about 90%; and with QZ bases, about 96−98%. However, there are still differences between the various construction schemes.

Again, the most accurate among the DZ bases is ano-pVDZ, which, on average, recovers about 4% or in absolute terms 20 mEh more correlation energy than the basis sets with uncontracted polarization functions. Here, the comparison with rcc-pVDZ and acc-pVDZ reveals that this is mainly due to the contracted polarization function and to a lesser extent to the more extensive s and p contractions. The differences among the other DZ bases are quite small.

A similar but much less pronounced result is obtained at the TZ level where ano-pVTZ is about 1−3% (or about 5 mEh on average) better than the other basis sets of the same size.

Interestingly, the differences almost vanish at the QZ level where there are only small differences between all QZ bases. All results are within 0.5% or 1−2 mEh of each other, and all deviate by 3−4% from the reference values.

It is obvious from the results obtained with sano-pVXZ+ and ano-pVXZ+ that limited improvements can be obtained from further augmentation of the basis set. The sano-pVDZ+ results are significantly better (~4%) than those obtained with ano-pVDZ, but already sano-pVTZ+ is rather similar to ano-pVTZ. The comparison between ano-pVXZ+ and ano-pV($n$+1)Z confirms that in order to reach the next level of accuracy it is more important to add the next higher angular momentum polarization function rather than to extend the existing polarization sets.

**4.3. Total Energies.** Since the ano-pVXZ basis sets of a given size have shown the best SCF and simultaneously the best correlation energies, it is trivial that they also show the best total energy in comparison with the reference values. The effects are most pronounced for the DZ and TZ basis sets where the ANO-basis sets are clearly superior to the "def2", "cc", and "pc" bases of the same size. For the QZ bases, all construction schemes start to converge to the same values, and the mean unsigned deviations are all between 13 and 16 mEh for the present test set. This error is dominated by the errors in the correlation energies that are still an order of magnitude larger than the SCF error.

**4.4. Explicitly Correlated (R12) Energies.** We also performed a series of CCSD(T)$_{R12}$ calculations using the newly developed ano-pVXZ basis sets as well as standard correlation-consistent basis sets (cc-pVXZ and aug-cc-pVXZ) and the recently developed cc-pVXZ-F12 basis set of Peterson et al., who specifically optimized their basis sets for R12 methods.[36] The immediate objective of these efforts was to examine whether the newly developed ANO basis sets are suitable for R12 methods, even though they were not constructed with such calculations in mind. We also wanted to see whether it is worth employing the ANO approach for constructing basis sets specifically suited for explicitly correlated (R12) methods.

Because the R12 correction only reduces the basis set error of the correlation energy, the basis set error of the SCF energy is relatively more significant in the context of explicitly correlated methods. The RMS error of the (aug)-cc-pVXZ SCF energy is larger than that of the correlation energy for X = D, is comparable to the latter for X = T, and is smaller than the latter for X = Q. With the ano-pVXZ basis sets—by virtue of their much improved SCF energy—the correlation energy error is always greater than the error of SCF energy. Peterson et al. used this observation in the design of the cc-pVXZ-F12 basis sets: the cc-pVXZ-F12 basis is similar in structure to the aug-cc-pVXZ basis, with the exception of more s and p primitives and the greater number of s and p shells in the former. The latter feature is largely responsible for the improved SCF energies obtained with the VXZ-F12 basis sets. The additional s and p functions dramatically reduce the basis set error of the cc-pV*X*Z-F12 SCF energy compared to its (aug)-cc-pVXZ counterpart. As a result, the basis set error of SCF energy with cc-pVXZ-F12 bases is always much smaller than that of the corresponding CCSD(T)$_{R12}$ correlation energy. The ano-pVXZ basis sets are comparable in their SCF energy performance to the cc-pVXZ-F12 counterparts, except for X = Q. The ano-pVXZ+ family, as expected, produces the smallest SCF energy errors.

Note that the error of SCF energy can be greatly reduced by including the CABS singles correction: indeed, the mean unsigned error of the total energy including the CABS correction is similar to the error of correlation energy with all basis sets. Although the evaluation of the CABS singles correction requires the evaluation of Coulomb integrals with two auxiliary basis set indices, its cost is negligible compared to that of the R12 correction. Without such a correction, the ano-pVXZ and cc-pVXZ-F12 basis sets produce smaller SCF energy errors and thus should be preferred to the cc-pVXZ and aug-cc-pVXZ series.

Let us now turn our attention to the basis set error of CCSD(T)$_{R12}$ correlation energy. The initial applications of modern R12 methods utilized the standard correlation-consistent basis set families. The cc-pVXZ series was, however, found to result in less accurate correlation energies than its aug-cc-pVXZ counterpart; thus all applications of R12 methods used the augmented correlation-consistent series. However, the aug-cc-pVXZ basis sets contain highly diffuse functions, and molecular studies in "real life" applications are often complicated by (near) linear dependencies in the basis set. For ANO basis sets, the higher members of a given angular momentum contain more and more nodes that extend into the outer region of the molecule, and hence these functions may take on the role of diffuse functions. However, the lack of uncontracted diffuse primitives in the ANO basis sets may lead to problems in the calculations of weak interactions and on anions or highly excited states. These subjects will be investigated in the future.

The performance of the newly developed ano-pVXZ series for R12 correlation energies falls in between the cc-pVXZ and aug-cc-pVXZ series: ano-pVDZ basis results in RMS correlation energy error similar to that of cc-pVDZ, whereas the ano-pVQZ basis is comparable in that sense to the aug-cc-pVQZ basis. The ano-pVXZ+ basis sets, which have the same number of functions as the aug-cc-pVXZ family, are significantly better than the latter. *These encouraging findings suggest that the uncontracted low-exponent (diffuse) basis functions do not have to be included in larger basis sets for R12 calculations*; as mentioned above, avoiding such functions is crucial for avoiding linear dependencies and poorly conditioned equations in computations on large molecules.

The ano-pVXZ basis sets are inferior to the cc-pVXZ-F12 basis sets for computing correlation energies with R12 methods. However, the ano-pVXZ+ basis sets, which have even slightly fewer basis functions than the cc-pVXZ-F12 basis sets, produce the most accurate correlation energies. For example, the RMS($E_C$) obtained with the ano-aug-pVQZ basis is 1.9 millihartree, which is 0.1 millihartree smaller than that with the cc-pVQZ-F12 basis. This is an interesting finding: the cc-pVXZ-F12 series was designed specifically for R12 methods, whereas the new ano-pVXZ+ series was not. For example, the structure of cc-pVXZ-F12 bases is similar to that of aug-cc-pVXZ; the former yield smaller RMS($E_C$)'s due to the reoptimized polarization functions;

this suggests that reoptimization of the (aug)-ano-pVXZ basis sets for R12 methods may improve their performance similarly.

Thus, the limited set of explicitly correlated calculations with the new ano-pVXZ+ basis set series hints that specific optimization of ANO basis sets for R12 calculations might be fruitful. Potential improvement to the level of what is obtained with the cc-pVXZ-F12 basis sets should be within reach. Such basis sets should be numerically amenable for applications to large molecules due to their lack of uncontracted diffuse functions (of course, diffuse functions will have to be included for computations on anions or Rydberg states).

## 5. Extrapolated Energies

One important aspect of the basis set construction is to investigate whether a given family of basis sets can be used to reliably extrapolate (components of) the energy to the basis set limit. Basis set extrapolation is a standard technique in atomic physics, where partial wave expansions suggest the appropriate extrapolation formula. In molecular calculations, basis set extrapolation is more empirical and relies on the systematic structure of the basis set series. Dunning's work on correlation consistent basis sets instantly spurred a flurry of developments in basis set extrapolation techniques. Several formulas for extrapolation of the correlation energy exist: the work of Helgaker et al., who suggested the use of inverse cubic extrapolation formula.[4,39] This formula suggests that the complete basis set limit correlation energy can be obtained as a linear combination of the correlation energies obtained with cc-basis sets $X$ and $Y$:

$$E_{\mathrm{corr}}^{\infty} = \frac{X^3 E_{\mathrm{corr}}^{(X)} - Y^3 E_{\mathrm{corr}}^{(Y)}}{X^3 - Y^3} \tag{1}$$

Truhlar pointed out that for practical computational chemistry applications, usually at most $X = 2$ and $Y = 3$ is feasible.[40] For these low cardinal numbers, the cc-bases do not reliably obey the asymptotic law, and hence Truhlar has investigated a more flexible form:

$$E_{\mathrm{corr}}^{\infty} = \frac{X^\beta E_{\mathrm{corr}}^{(X)} - Y^\beta E_{\mathrm{corr}}^{(Y)}}{X^\beta - Y^\beta} \tag{2}$$

and determined the optimum exponent $\beta$ to be 2.4 for the combination of cc-pVDZ, cc-pVTZ, and CCSD or CCSD(T) from calculations on Ne, HF, and $H_2O$. Truhlar's approach was later shown by Schwenke[41] to be a special case of a general approach that approximates the CBS limit energy as a linear combination of energies computed with $X$, $Y$, $Z$, etc. basis sets; coefficients in the linear expansion are specific to each method (SCF, CCSD, etc.) and each set of basis sets. For two basis sets, eq 2 can be rewritten in Schwenke's form:[41]

$$E_{\mathrm{corr}}^{(\cdot)} = \frac{X^\beta E_{\mathrm{corr}}^{(X)} - Y^\beta E_{\mathrm{corr}}^{(Y)}}{X^\beta - Y^\beta} = \frac{X^\beta E_{\mathrm{corr}}^{(X)} - Y^\beta E_{\mathrm{corr}}^{(X)} + Y^\beta (E_{\mathrm{corr}}^{(X)} - E_{\mathrm{corr}}^{(Y)})}{X^\beta - Y^\beta}$$
$$= E_{\mathrm{corr}}^{(X)} + \frac{Y^\beta}{X^\beta - Y^\beta}(E_{\mathrm{corr}}^{(X)} - E_{\mathrm{corr}}^{(Y)}) = E_{\mathrm{corr}}^{(X)} + f(X,Y)(E_{\mathrm{corr}}^{(X)} - E_{\mathrm{corr}}^{(Y)}) \tag{3}$$

Hence there is a 1:1 correspondence between exponent $\beta$ and the unknown linear coefficient $f(X,Y)$. We have followed this recipe by using eq 2 and reoptimized the exponent $\beta$ for each pair of basis sets that was investigated.

Perhaps surprisingly, the extrapolation of the SCF energy appears to be more difficult than the extrapolation of the correlation energy. Halkier et al.[43] established that exponential extrapolation—as long pursued by Feller[44,45]—is more successful than the power law used by Truhlar. More recently, Petersson and co-workers have studied the problem in detail and have used[46] an extrapolation of the form:

$$E_{\mathrm{SCF}}^{(X)} = E_{\mathrm{SCF}}^{(\infty)} + A \exp(-\alpha\sqrt{X}) \tag{4}$$

(the equation is attributed to Karton and Martin[47]). For their new basis sets (*n*ZaP), the constant $\alpha = 6.3$ was found to be universally valid. We have followed Petersson and co-workers but have fitted $\alpha$ to each pair of basis sets used for extrapolation.

Thus, in the extrapolation study, we have fitted one parameter to the SCF energies and one parameter to the correlation energies to minimize the total mean unsigned error for the 21-molecule test set; the optimal values of parameters are given in Table 3. Qualitatively, extrapolations that lead to $\beta = 3$ and a relatively small $\alpha$ are preferred. Under these circumstances, the correlation energy extrapolation is thought to be most physically sound, and the less steep SCF energy extrapolation is expected to be more robust in actual applications.

Among the 2/3 extrapolations, the rcc-pV(D/T)Z pair slightly outperforms the other basis sets of the same size, and only cc-pV(D/T)Z comes close. Hence, the use of contracted polarization functions does appear to improve the quality of at least the lower member of the cc-basis set family. Nevertheless, the differences between the various basis sets are not large and amount to about 0.5 mEh in the total energy on average. Augmentation of the ANO basis sets does improve the results, and with ano-pV(D/T)Z+, the mean unsigned error of the extrapolated total energy of 1.27 mEh must be considered as an excellent result. It is, in our opinion, impressive that with such a simple extrapolation scheme one can compute the reference energy with an average unsigned error of less than 2 mEh. In particular, the error in the SCF energy can be reduced to below 1 mEh using the extrapolation formula of Petersson and co-workers, while the Truhlar−Schwenke extrapolation provides correlation energies within 2−3 mEh from the reference values. The results lend credence to Truhlar's choice of $\beta = 2.46$. All extrapolations with the 2/3 pairs come close to this value.

All 3/4 extrapolations reduce the SCF error to 0.1−0.4 mEh, with the best results being obtained with the def2 basis sets. However, pc-*n*, ano-pVXZ, sano-pVXZ+, rcc-pVXZ, and cc-pVXZ are all similarly good. The extrapolation of

***Table 3.*** Errors of the Extrapolated SCF and Correlation Energies Relative to the Reference Data (in mEh where Applicable)

| | 2/3 extrapolation | | | | | 3/4 extrapolation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha_{23}$ | $\beta_{23}$ | MUE($E_{HF}$) | MUE($E_C$) | MUE($E_{tot}$) | $\alpha_{34}$ | $\beta_{34}$ | MUE($E_{HF}$) | MUE($E_C$) | MUE($E_{tot}$) |
| cc-pVXZ | 4.42 | 2.46 | 0.61 | 2.00 | 1.77 | 5.46 | 3.05 | 0.22 | 0.60 | 0.75 |
| aug-cc-pVXZ | 4.30 | 2.51 | 0.65 | 2.06 | 1.86 | 5.79 | 3.05 | 0.17 | 0.53 | 0.68 |
| pc-*n* | 7.02 | 2.01 | 0.90 | 2.69 | 2.76 | 9.78 | 4.09 | 0.17 | 0.41 | 0.28 |
| def2 | 10.39 | 2.40 | 0.47 | 2.10 | 2.01 | 7.88 | 2.97 | 0.10 | 0.66 | 0.67 |
| ano-pVXZ | 5.41 | 2.43 | 0.72 | 2.84 | 2.20 | 4.48 | 2.97 | 0.18 | 0.62 | 0.64 |
| sano-pVXZ+ | 5.48 | 2.21 | 0.32 | 1.86 | 1.68 | 4.18 | 2.83 | 0.11 | 0.28 | 0.32 |
| ano-pVXZ+ | 5.12 | 2.41 | 0.20 | 1.29 | 1.27 | 5.00 | 2.52 | 0.09 | 0.31 | 0.38 |
| acc-pVXZ | 4.80 | 2.34 | 0.85 | 1.61 | 1.81 | 4.92 | 2.94 | 0.36 | 1.17 | 1.29 |
| rcc-pVXZ | 4.43 | 2.47 | 0.69 | 1.82 | 1.44 | 5.46 | 3.00 | 0.17 | 0.56 | 0.72 |
| Roos-ANO | 5.15 | 2.50 | 0.41 | 2.19 | 2.22 | | | | | |

the Roos-ANO with the NASA-AMES set is not fair and has therefore been omitted from Table 3. Pleasingly, all 3/4 extrapolations (except the one for pc-*n*) come up with an optimum $\beta$ that is close to the preferred value of 3. Hence, for all intents and purposes, it could be replaced by just 3 once one proceeds beyond the 2/3 extrapolation level. In any event, the extrapolated 3/4 energies almost all fall within 1 mEh of the reference values, which is considered pleasing. From this perspective, the choice of the particular 3/4 set might not be overly important and could be done under consideration of computational convenience. In particular, the results demonstrate that the def2-TZVPP/def2-QZVPP pair is well suitable for extrapolation, and the pc-2/pc-3 result is even one of the best in the test set. We do, however, view the latter result with some reservation due to the unphysical exponent of 4.09 in the correlation energy extrapolation, the steep SCF exponent, and a large amount of error cancellation between the errors in the SCF and correlation energy extrapolations. In terms of the SCF extrapolation, the ANO basis sets lead to the least steep extrapolation and may therefore be preferred.

## 6. Comparison of Extrapolation and Explicit Correlation

Despite the impressive improvement of correlation energy due to the use of explicitly correlated methods, the total R12 energies obtained with ano-pVTZ and aug-cc-pVTZ basis sets are approximately 3.5 times less accurate than the corresponding 2/3 extrapolated values. The corresponding quadruple-$\zeta$ ratio is approximately 4. However, at this level, the errors are so small that they probably fall within the error bar of the reference data itself that was "only" obtained through 5/6 extrapolation. The use of cc-pVXZ-F12 basis sets optimized specifically for explicitly correlated methods improves the comparison in favor of R12 methods, but nevertheless, extrapolated energies are more accurate by a factor of 3. The comparison with ano-pVXZ+ is most favorable to the R12 calculations: they are only a factor of 2−2.5 worse than the extrapolated results. Thus, at present, the R12 methods are not competitive with extrapolation for the total energies. However, one can, of course, combine R12 methods with extrapolation, as already pursued by Hill et al.[48]

It would be imprudent to overly generalize these findings. Our past experience with extrapolation methods suggests that their spectacular performance for total energies does not transfer intact when applied to relative energies.[48] This is particularly evident from the fact that the extrapolated correlation energies may either overshoot or undershoot the true correlation energy, while the error in the correlation energies is always positive in the explicitly correlated methods. Although the Hylleraas functional for the R12 correlation energy provides an upper bound to the exact value, in practice, the upper bound property is not guaranteed due to the inexact Hartree−Fock solution and various approximations involved in R12 methods. The violations are, however, small, and with the modern R12 methods, the basis set limit of energy is approached from above (the basis set error is positive). A more complete and fair comparison of the extrapolation methods vs R12 methods will involve the construction of ANO-type basis sets specifically optimized for R12 methods, and application to relative energies of interest to chemists (reaction energies, activation barriers). This study is outside the scope of this work.

## 7. Computational Details

All calculations reported in this work were carried out with a development version of the ORCA program package;[49] the MPQC (http://www.mpqc.org/mpqc-snapshot-html/index.html, accessed July 11th, 2010) and Psi3 (http://www.psicode.org/index.html, accessed July 11th, 2010) packages were used for all explicitly correlated calculations. Geometries were optimized with the B3LYP functional[50−52] in conjunction with the def2-TZVP[53] basis set. The geometrical parameters thus obtained are (atomic units and degrees): $H_2$ ($R$ = 1.406421), $BH_3$ ($R_{BH}$ = 1.1898, $D_{3\,h}$), $CH_4$ ($R_{CH}$ = 2.05898, $T_d$), $NH_3$ ($R_{NH}$ = 1.91599, H−N−$H$ = 107.312°, H−N−H−$H$ = 115.067°), $H_2O$ ($R_{OH}$ = 1.81975, H−O−$H$ = 105.237°), FH ($R$ = 1.747475), $B_2$ ($R$ = 3.661983), BC ($R$ = 2.554910), BN ($R$ = 2.491487), BO ($R$ = 2.272142), BF ($R$ = 2.390700), $C_2$ ($R$ = 2.357244), CN ($R$ = 2.319042), CO ($R$ = 2.126086), CF ($R$ = 2.410890), $N_2$ (2.061850), NO ($R$ = 2.164009), NF ($E$ = 2.491128), $O_2$ ($R$ = 2.276277), OF ($R$ = 2.550064), $F_2$ ($R$ = 2.639683). Coupled cluster CCSD(T) calculations were performed with the 1s orbital kept frozen (except for hydrogen, of course). Open shell species were treated in the unrestricted formalism and without restrictions on the spatial part of the wave function.

Explicitly correlated CCSD(T) computations on closed-shell molecules were performed with the CCSD(T)$_{R12}$ method of Valeev et al.[26,54,55] The geminal amplitudes were fixed at the values determined by the first-order cusp conditions.

Three- and four-electron integrals were approximated using the CABS+ approach[56] with the uncontracted cc-pV6Z basis as the auxiliary basis set. The standard R12/C approximation[57] was used throughout to approximate the so-called B intermediate of the R12 theory. Exponential (Slater-type) correlation factor $e^{-\gamma r_{12}}$ fitted to six Gaussian geminals was used in all calculations; geminal exponent $\gamma = 1.4\ a_0$ was used in all calculations except those involving the cc-pVXZ-F12 basis sets,[58] for which the recommended geminal exponents were used. The R12 correction of the CCSD(T)$_{R12}$ method was computed using robust density fitting (DF);[42] the following {orbital, density-fitting} basis set combinations were used: {ano-pVXZ, aug-cc-pV6Z-RI (C. Hättig, University of Bochum, Germany, unpublished)}, {cc-pVXZ, cc-pV(X+1)Z-RI},[59,60] {aug-cc-pVXZ, aug-cc-pV(X+1)Z-RI}, and {cc-pVXZ-F12, cc-pV(X+1)Z-RI}. Total energies reported in the rightmost column of Table 2 also include the second-order energy correction from single excitations to the CABS manifold correction.[26]

Atomic calculations were performed with uncontracted Gaussian basis sets as described in the main body of the paper. Calculations were tightly converged on the atomic ground states: He($^1$S), B($^2$P), C($^3$P), N($^4$S), O($^3$P), F($^2$P), and Ne ($^1$S). MR-ACPF calculations were performed on top of the CASSCF states using all configurations in the CAS including those of the wrong symmetry or with an insufficient number of open-shell orbitals as references. No contraction or selection of the interacting space was employed, and the calculations were converged to $10^{-12}$ Eh. The first order density was constructed and diagonalized in order to obtain atomic natural orbital contractions.

## 8. Conclusions

In this work, basis set construction strategies for correlated calculations on molecules were investigated. In particular, it was studied whether the family of correlation consistent basis sets could possibly be improved by incorporating elements of atomic natural orbital construction in their design. The answer to this question is positive: ANO basis sets deliver—provided they are derived from a sufficiently accurate set of primitive Gaussians—much better SCF and slightly better correlation energies than correlation consistent basis sets. The replacement of the uncontracted polarization functions with the contracted ANO ones (rcc-pVXZ) yields a significant but still somewhat limited improvement in the calculated correlation energies, while replacing the s and p parts of the cc-bases with their ANO counterparts (acc-pVXZ) does improve the SCF energies but not the correlation energy. Thus, one seems to be well advised to keep the full ANO contractions in order to obtain the best results at a given basis set size. This also is the most consistent strategy if algorithms can be devised to obtain the integrals over generally contracted basis sets efficiently for larger molecules. This will be discussed elsewhere.

A significant finding of the present study is the large increase in the accuracy of the SCF energies if one only supplements the ano-pVXZ with the next set of valence functions (saug-pVXZ). The increase in the number of basis functions by four per heavy atom is already limited at the DZ level (19 instead of 15 basis functions) and very small for the higher members of the basis set family. Since the integrals over all primitive members of the underlying set must be calculated anyway, the increase in computation time cannot be overly disturbing once the computational procedure is highly optimized. The full augmentation with the next set of polarization functions (ano-pVXZ+) comes at a high computational cost in the correlation energy calculation, while the gains in accuracy appear to be somewhat limited compared to the calculation with the next higher cardinal number.

The gain in accuracy obtained with the ano-pVXZ basis sets relative to their cc-pVXZ parents is considerable at the double-$\zeta$ level, limited but significant at the triple-$\zeta$ level, and more limited for the quadruple-$\zeta$ variant. Indeed, it was particularly the performance of the double- and triple-$\zeta$ cc bases that was the main motivation for the present work. With explicitly correlated R12 methods, the performance of ano-pVXZ+ bases was superior to that of the cc-pVXZ-F12 basis sets of Peterson and co-workers that were specifically optimized for R12 methods. The combination of these compact basis sets lacking numerically troublesome uncontracted diffuse basis functions with explicit correlation and local correlation approaches appears to be a fruitful route toward accurate computational chemistry. The results of Werner and co-workers[61−63] along these lines appear to be very promising.

The situation changes somewhat if the SCF and correlation energies are extrapolated to the basis set limit. In this work, the SCF extrapolation technique recently proposed by Peterson and co-workers has been combined with the correlation energy extrapolation of Truhlar that is based on the work of Helgaker and co-workers as well as Schwenke.[64] One parameter was optimized for the SCF energies and one for the CCSD(T) correlation energy for each pair of basis sets considered. Among the double-/triple-$\zeta$ basis set combinations, the cc-pV(D/T)Z sets were again found to perform best, but all extrapolations lead to total energies within 2.5 mEh of the reference values. Again, augmentation improves the results obtained with the ANO bases. For the triple−quadruple-$\zeta$ basis set extrapolations, the differences between the various basis set combinations become rather small, and one typically reaches better than 1 mEh deviation from the reference values. It is noteworthy that even with basis sets that were optimized for SCF calculations (pc-2/pc-3, def2-TZVPP/def2-QZVPP) one obtains excellent extrapolations to the SCF and CCSD(T) correlation basis set limits. While for the double/triple-$\zeta$ basis set pairs Truhlar's observation[39] that the optimum extrapolation coefficient for the correlation energy is around 2.4 was confirmed, for the triple/quadruple-$\zeta$ basis set pairs, all basis set pairs (except pc-2/pc-3) show optimized exponents close to the theoretically expected values of 3, which is considered pleasing. Recently, Petersson and co-workers have reported an effort toward extrapolation techniques for molecular calculations and have devised a new series of basis sets. So far, they have published results for SCF energies[40] and more recently also for MP2.[46] It will be interesting to compare these results with the ones that have been obtained in this work once they become publically available.

Having realized the superior accuracy of ano-pVXZ sets compared to segmented bases, it is necessary to develop computational algorithms that ease the burden of integral evaluation. Most integral programs take little advantage of contraction, and the cost of integral evaluation increases with the fourth power of the contraction depth. This clearly leads to extremely expensive calculations for the deeply contracted ano-pVXZ basis sets. Thus, unless procedures that are specifically targeted toward ANO contraction are employed, the use of the ano-pVXZ or any other ANO basis in routine molecular calculations is not attractive. Such techniques have been developed in the framework of the ORCA program and will be reported elsewhere. For a long time, the MOLCAS package has also been optimized for this task, as is amply documented by the many successful molecular applications on the basis of the Lund group ANO basis sets. We do not claim great superiority of our ANO sets over the ones initially constructed by Almlöf and Taylor[19−21] or by Roos and co-workers.[21,22,65] The emphasis of the present work was the comparison between various basis set construction strategies with the main result that the ANO construction scheme appears to be the most successful one if only the number of basis functions but not the contraction depth is of concern. An efficient implementation of an ANO-based SCF and coupled cluster program within the ORCA package will be described elsewhere.

The various ANO basis sets discussed in this paper have been constructed from H−Ar and will be provided to the public via the EMSL basis set library and the Supporting Information of this manuscript. We expect that for first row transition metals, the basis sets designed by Roos and co-workers can safely be combined with the ano-pVXZ series.

**Acknowledgment.** F.N. gratefully acknowledges financial support of this work by the special research units SFB 813 ("Chemistry at Spin Centers") and SFB 624 ("Template Chemistry"), both centered at the University of Bonn. E.F.V. is grateful to the Donors of the American Chemical Society Petroleum Research Fund (Grant No. 46811-G6) and the U.S. National Science Foundation (CAREER Award No. CHE-0847295 and CRIF:MU Award No. CHE-0741927). E.F.V. is an Alfred P. Sloan Research Fellow and a Camille Dreyfus Teacher−Scholar.

**Supporting Information Available:** Listing of the proposed basis sets. This material is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Kutzelnigg, W.; Morgan, J. D. *J. Chem. Phys.* **1992**, *96*, 4484–4508.

(2) Kutzelnigg, W. *Theor. Chim. Acta* **1985**, *68*, 445–469.

(3) Helgaker, T.; Klopper, W.; Tew, D. P. *Mol. Phys.* **2008**, *106*, 2107–2143.

(4) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

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

(6) Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

(7) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

(8) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **1999**, *110*, 7667–7676.

(9) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1995**, *103*, 4572–4585.

(10) Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 5829.

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

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

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

(14) Dunning, J. T. H. *J. Chem. Phys.* **1980**, *90*, 1007.

(15) Wilson, A. K.; van Mourik, T.; Dunning, T. H. *THEOCHEM* **1996**, *388*, 339–349.

(16) Peterson, K. A.; Wilson, A. K.; Woon, D. E.; Dunning, T. H. *Theor. Chem. Acc.* **1997**, *97*, 251–259.

(17) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.

(18) Figgen, D.; Peterson, K. A.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2009**, 130.

(19) Figgen, D.; Peterson, K. A.; Stoll, H. *J. Chem. Phys.* **2008**, *128*.

(20) Peterson, K. A.; Figgen, D.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2007**, 126.

(21) Pouamerigo, R.; Merchan, M.; Nebot-Gil, I.; Widmark, P. O.; Roos, B. O. *Theo. Chim. Acta* **1995**, *92*, 149–181.

(22) Pierloot, K.; Dumez, B.; Widmark, P. O.; Roos, B. O. *Theor. Chim. Acta* **1995**, *90*, 87–114.

(23) Widmark, P. O.; Joakim, B.; Roos, B. O. *Theor. Chim. Acta* **1991**, *79*, 419–432.

(24) Widmark, P. O.; Malmqvist, P. A.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.

(25) Dunlap, B. I. *THEOCHEM* **2000**, *529*, 37–40.

(26) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(27) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(28) Jensen, F. *Theor. Chem. Acc.* **2005**, *113*, 267–273.

(29) Jensen, F.; Helgaker, T. *J. Chem. Phys.* **2004**, *121*, 3463–3470.

(30) Jensen, F. *J. Chem. Phys.* **2003**, *118*, 2459–2463.

(31) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 3502–3502.

(32) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 7372–7379.

(33) Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234–9240.

(34) Jensen, F. *J. Chem. Phys.* **2001**, *115*, 9113–9125.

(35) Gdanitz, R. J.; Ahlrichs, R. *Chem. Phys. Lett.* **1988**, *143*, 413–420.

(36) Almlöf, J.; Taylor, P. R. *Adv. Quantum Chem.* **1991**, *22*, 301–373.

(37) Almlöf, J.; Taylor, P. R. *J. Chem. Phys.* **1987**, *86*, 4070–4077.

(38) Peterson, K. A.; Adler, T. B.; Werner, H. J. *J. Chem. Phys.* **2008**, *128*, 084102.

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

(40) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.

(41) Schwenke, D. W. *J. Chem. Phys.* **2005**, *122*, 014107.

(42) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H. J. *J. Chem. Phys.* **2009**, *131*, 194105.

(43) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.

(44) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104.

(45) Feller, D. *J. Chem. Phys.* **1993**, *98*, 7059.

(46) Zhong, S. J.; Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2008**, *129*.

(47) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330–333.

(48) Hill, J. G.; Mazumder, S.; Peterson, K. A. *J. Chem. Phys.* **2010**, 132.

(49) Valeev, E. F.; Crawford, T. D. *J. Chem. Phys.* **2008**, *128*, 244113.

(50) Neese, F.; Becker, U.; Ganyushin, D.; Kollmar, C.; Kossmann, S.; Hansen, A.; Liakos, D.; Petrenko, T.; Riplinger, C.; Wennmohs, F. ORCA - an ab initio, density functional and semiempirical program package, version 2.8.0; University of Bonn: Bonn, Germany, 2010.

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

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

(53) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(54) Valeev, E. F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 106–113.

(55) Torheyden, M.; Valeev, E. F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3410–3420.

(56) Valeev, E. F.; Crawford, T. D. *J. Chem. Phys.* **2008**, 128.

(57) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190–195.

(58) Kedzuch, S.; Milko, M. J. N. *Int. J. Quantum Chem.* **2005**, *105*, 929–936.

(59) Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 4607–4613.

(60) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175.

(61) Adler, T. B.; Knizia, G.; Werner, H. *J. Chem. Phys.* **2007**, *127*, 221106.

(62) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.

(63) Adler, T. B.; Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2009**, *130*, 054106.

(64) Werner, H.-J. *J. Chem. Phys.* **2008**, *129*, 101103.

(65) Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2010**, *132*, 114111.

CT100396Y

JCTC Journal of Chemical Theory and Computation

# Modeling Charge Resonance in Cationic Molecular Clusters: Combining DFT-Tight Binding with Configuration Interaction

Mathias Rapacioli,* Fernand Spiegelman, Anthony Scemama, and André Mirtschink

*Université de Toulouse, UPS, LCPQ (Laboratoire de Chimie et Physique Quantiques), IRSAMC, 118 Route de Narbonne, F-31062 Toulouse, France, and CNRS, LCPQ (Laboratoire de Chimie et Physique Quantiques), IRSAMC, F-31062 Toulouse, France*

Received July 23, 2010

**Abstract:** In order to investigate charge resonance situations in molecular complexes, Wu et al. (*J. Chem. Phys.* **2007**, *127*, 164119) recently proposed a configuration interaction method with a valence bond-like multiconfigurational basis obtained from constrained DFT calculations. We adapt this method to the Self-Consistent Charge Density-Functional-based Tight Binding (SCC-DFTB) approach and provide expressions for the gradients of the energy with respect to the nuclear coordinates. It is shown that the method corrects the wrong SCC-DFTB behavior of the potential energy surface in the dissociation regions. This scheme is applied to determine the structural and stability properties of positively charged molecular dimers with full structural optimization, namely, the benzene dimer cation and the water dimer cation. The method yields binding energies in good agreement with experimental data and high-level reference calculations.

## 1. Introduction

The description of neutral molecular clusters requires the consideration of various contributions of the intermolecular energy, including Pauli repulsion, polarization, electrostatics (static multipole interactions), induction forces (multipole-induced multipole interactions), and London dispersion. The treatment of the electronic structure of singly ionized molecular clusters also needs to consider charge resonance, which may cause the charge to be partially or totally delocalized over the molecular units and polarization contributions due to the influence of the charge. Both lead to a stabilization of the charged species as compared to the analogous neutrals. A proper description requires correct balance between charge delocalization and polarization forces.

While Density Functional Theory (DFT) is an appealing method for describing the electronic properties of clusters with dozens, maybe hundreds, of atoms, at least in single point calculations, most common functionals are known to fail in properly describing dispersion forces. This is the first handicap to deal with by treating molecular clusters. The search for new functionals accounting for dispersion[1−9] (for a review, see ref 10) is a very active field, while semiempirical corrections to standard DFT calculations are also used.[11−18] The description of charge resonance in molecular clusters is another serious problem in standard density functional approaches. Using the Kohn−Sham formalism with these functionals, one arising problem is due to the self-interaction of the delocalized charge. Many investigations have addressed the analysis and correction of the self-interaction error arising with approximated DFT functionals (see, for instance, refs 19−38). This error is particularly prevalent in the dissociation of radical cations.[35,36] Similarly, a cationic molecular dimer involving two identical units should dissociate into one molecular cation and one neutral, but in a restricted DFT scheme, the charge is asymptotically equally shared by the two units, breaking the energy additivity and further introducing a spurious Coulomb interaction between the two moieties. Although such an artifact is essential in the dissociation, it is also expected to play a role all over the potential energy surface, including the equilibrium geometries.

* To whom correspondence should be addressed. E-mail: mathias.rapacioli@irsamc.ups-tlse.fr.

10.1021/ct100412f 
Published on Web 12/13/2010

A correct description of dissociation is in principle easily obtained by the use of a multiconfigurational wave function. It can be achieved by high-level methods like Configuration Interaction[39,40] (CI)-based methods (Multi-Configurational Self-Consistent Field,[41] MCSCF; Multi-Reference Configuration Interaction,[42] MRCI) or Coupled Cluster[43] (CC) approaches (for a review, see ref 44) but at a high computational cost. Such calculations may provide benchmarks on reasonably small systems (essentially dimers) but rapidly exceed today's possibilities as soon as the molecular units exceed a few tens of atoms.

One of the tracks for circumventing the drawbacks of the present state DFT in an *ab initio* framework is to combine CI, for describing long-range (lr) electron−electron interactions, and DFT, for the short-range electron−electron interactions (sr). This gave rise to the lr−sr formalism following Savin's formulation,[45−48] enabling combinations of Møller−Plesset (MP) perturbation,[49−51] CC, and/or CI approaches with DFT. This formulation is quite attractive; nevertheless, its numerical cost is significantly larger than that of a standard DFT calculation.

Alternatively, charge resonance (or excitation resonance) is described quite simply in valence bond-like approaches[52−55] by explicitly considering the multiconfigurational nature of the wave function via the definition of a basis arising from configurations in which the charge (or excitation) is localized on a given fragment of the system. This is the essence of the excitonic models originating from solid state physics (see for instance ref 56 and references therein) but also used in molecular materials and even biological systems. An application to cationic molecular clusters of polycyclic aromatic hydrocarbon (PAH) was published by Bouvier et al.,[57] defining a resonance charge model based on frozen molecules and parametrized from *ab initio* CI calculations on dimers. Diatomics-in-molecule modeling of singly ionized rare gas clusters can also be expressed in a valence bond picture with a basis of atom-localized hole configurations and no internal geometrical structure.[58−60] That paved the way for extensive simulations of the electronic and dynamical properties of ionized rare gas clusters (see for instance Calvo et al.[61,62]).

More recently, the concept of a valence bond configuration description in a DFT framework was proposed by the group of Van Voorhis et al.[63−69] to investigate charge delocalization in mixed valence compounds exhibiting possible bistability with the perspective of controlling charge transfer. They developed a method combining Constrained DFT,[63−65] used to build charge localized configurations, with a small CI-like scheme (CDFT-CI) to deal with charge delocalization in extended systems. From the computational point of view, it is extremely appealing for singly charged clusters, since the static correlation associated with charge resonance is treated by the CI-like scheme, which in this case is linear scaling (as would be complete active space self-consistent field CASSCF[70] with a single hole in the MOs resulting from the HOMOs of the individual molecules), and the dynamical correlation is treated at the DFT level in a single configuration scheme (whereas a CASSCF would need complementing with dynamical correlation, for instance, CASPT2[71]). For addressing large systems, Self-Consistent-Charge Density-Functional-based Tight Binding (SCC-DFTB)[72−75] is interesting since it is computationally faster than DFT. It is derived from DFT through several approximations allowing the use of tabulated overlap and interaction integrals.

As SCC-DFTB is derived from DFT, it also inherits its lack of describing charge resonance with standard functionals essentially due to the self-interaction error. Detailed analysis of this problem in DFT proposals for self-interaction free functionals were given by Grafenstein et al.[35,36,76] Interestingly, some of those schemes produce localized orbitals. The transfer of such a concept within the DFTB framework would certainly be of interest. However, self-interaction corrections should introduce many centers' contributions into the DFTB parameters representing the Coulomb-exchange-correlation contribution, beyond the two-center approximations for electron−electron interaction integrals, a key point of the DFTB efficiency. This would require analytical assumptions for these terms, further parametrization, and transferability checking.

Recently, we presented a preliminary application of the CDFT-CI method in the SCC-DFTB framework also using approximations to determine the CI couplings. We studied coronene clusters with constrained geometries, because of the lack of gradients.[77] One of the interests of CDFTB-CI is to be safe in regard to all dissociation channels, even multicenter fragmentation.

In the present paper, we present the general adaptation of the CDFT+CI method to the SCC-DFTB framework, with the aim of future investigations of charge resonance in molecular clusters with large sizes. This method is called DFTB-VBCI (Valence Bond CI). In order to perform geometry optimizations, we also derive analytical expressions for the energy gradients with respect to the nuclear coordinates. It allows us to achieve full structural optimization for the benzene dimer and water dimer cations, respectively.

Section 2 is devoted to the presentation of the general methodology, the DFTB-VBCI approach, and the derivation of analytical expressions for the nuclear forces. In section 3, we benchmark the method on ionic benzene and water dimers on the basis of comparisons with high-level calculations. A summary and perspectives are given in section 4.

## 2. Methodology

The DFTB-VBCI method is an adaptation of the CDFT+CI approach[63−69] to the SCC-DFTB scheme with the aim of treating charge resonance in ionized molecular clusters. In this approach, the wave function of the system $\Psi$ is expressed in a basis $\{\Phi^I\}$ of configurations. For each configuration, the charge is localized on a given fragment of the system. The intuitive decomposition of a molecular cluster leads to identifying the $N_{\mathrm{frag}}$ fragments as the monomers, and the wave function becomes

$$\Psi = \sum_{I}^{N_{\mathrm{frag}}} b_I \Phi^I \tag{1}$$

where $\Phi^I$ is the configuration where the charge is fully carried by fragment $I$. Each charge localized configuration $\Phi^I$ is a single Slater determinant, built from the molecular

orbitals (MO) $\{\phi_i^I\}$ resulting from a constrained SCC-DFTB calculation. These VB like configurations then interact within a small CI-like scheme, giving their coefficients $b_I$ in the wave function and the ground state energy.

In this methodological part, we first briefly recall the SCC-DFTB scheme basics (section 2.1) before explaining the derivation of the charge localized configurations $\Phi^I$ using the constrained SCC-DFTB (section 2.2) and the CI-like scheme calculation (section 2.3). We present then analytical expressions for the nuclear gradients (section 2.4) and some further approximations to accelerate the approach (section 2.5). We adopt different font conventions to distinguish between matrices expressed in different basis sets. For instance, a Hamiltonian matrix is written as $H$ in the atomic orbital (AO) basis set; $\boldsymbol{H}$ in the MO basis set, and $\mathbf{H}$ in the determinant basis set (the basis of the charge-localized configurations).

**2.1. DFT and SCC-DFTB.** Several reviews on the DFTB and SCC-DFTB methods can be found in the literature.[72−75] SCC-DFTB differs from Kohn−Sham DFT expressed on a local basis set according to the following approximations: (i) The DFT energy is expanded up to the second order with respect to charge density fluctuations around a given reference density. (ii) All three center interaction integrals are neglected as well as two center integrals involving atomic orbitals belonging to the same atom. (iii) The MOs are expressed in a minimal atomic basis set

$$\phi_i = \sum_\mu c_{i\mu}\varphi_\mu \tag{2}$$

(iv) The short distance repulsive potential is expressed as a function of two body interactions. (v) The second-order term in the DFT energy expansion is expressed as a function of atomic Mulliken charges and a $\Gamma$ matrix. With those approximations, the total SCC-DFTB energy reads

$$E^{\text{SCC-DFTB}} = \sum_{\alpha,\beta\neq\alpha}^{\text{atoms}} E_{\alpha\beta}^{\text{rep}} + \sum_i n_i\langle\phi_i|\hat{H}^0|\phi_i\rangle + \frac{1}{2}\sum_{\alpha,\beta}^{\text{atoms}} \Gamma_{\alpha\beta}q_\alpha q_\beta \tag{3}$$

where $\hat{H}^0$ is the Kohn−Sham operator at the reference density and $E_{\alpha\beta}^{\text{rep}}$ is the repulsive potential between atoms $\alpha$ and $\beta$. The matrix elements of $\hat{H}^0$ expressed in the atomic basis set as well as $\Gamma_{\alpha\beta}$ and $E_{\alpha\beta}^{\text{rep}}$ are interpolated from two body DFT calculations. $n_i$ represents the atomic orbital occupation numbers, and $q_\alpha$ represents the atomic Mulliken charges. The energy minimization is obtained by self-consistently solving the secular equation

$$\sum_\nu c_{i\nu}(H_{\mu\nu} - \varepsilon_i S_{\mu\nu}) = 0 \quad \forall\mu, i \tag{4}$$

$S$ is the atomic basis overlap matrix, and the Hamiltonian matrix reads $H = H^0 + H^1$ with

$$H^1_{\mu\in\alpha;\nu\in\beta} = \frac{1}{2}S_{\mu\nu}\sum_\xi^{\text{atoms}}(\Gamma_{\alpha\xi} + \Gamma_{\xi\beta})q_\xi \tag{5}$$

where $\mu \in \alpha$ means that the atomic orbital $\mu$ belongs to atom $\alpha$.

Additional terms can be added to account for London dispersion ($E^{\text{disp}}$) forces as a sum over atomic pairs.[14,78,79] The deMonNano code[80] was used as a starting point to implement these developments.

**2.2. Constrained SCC-DFTB.** Similarly to the constrained DFT,[63−65] the MOs $\{\phi_i^I\}$, used to build the configuration $\Phi^I$, are obtained from a minimization of the SCC-DFTB energy with the constraints that the charge is carried by fragment $I$ and that the orbitals are orthonormalized. The corresponding Lagrangian is

$$\mathcal{L} = E^{\text{SCC-DFTB}}(\{\phi_i^I\}) + \sum_{ij}\Lambda_{ij}^I(\langle\phi_i^I|\phi_j^I\rangle - \delta_{ij}) + V^I\left(\sum_i n_i\langle\phi_i^I|\hat{P}^I|\phi_i^I\rangle - N^I\right) \tag{6}$$

where $V^I$ is the Lagrange multiplier ensuring the charge localization constraint, $\hat{P}^I$ is the projector of the density on fragment $I$, $N^I$ is the number of electrons on fragment $I$, which constrains the charge to be localized on this fragment, and $\Lambda_{ij}$ represents the Lagrange multipliers ensuring the orbitals' orthonormality constraints. Wu and Van Voorhis[65] discussed the effect of several localization schemes, based on different charge definitions (Mulliken,[81] Löwdin,[82] and Becke's multicenter integration scheme[83]), on the constrained energy, finally using the Löwdin approach. We used for the constrained SCC-DFTB the Mulliken charge definition because (i) the defects of Mulliken charges are less crucial in SCC-DFTB than in DFT due to the use of a minimal atomic basis set (no diffuse functions) and, (ii) in the most used version, SCC-DFTB is a Mulliken charge-based approach, and all of the matrices have been parametrized for this charge definition. This choice leads to the expression for the constraint

$$\sum_{i\nu\mu} n_i c_{i\nu}^I c_{i\mu}^I P_{\nu\mu}^I = N^I \tag{7}$$

with $P^I$ being the projection matrix expressed as[65]

$$P_{\mu\nu}^I = \begin{cases} 0 & \text{if } \mu \notin I \text{ and } \nu \notin I \\ S_{\mu\nu} & \text{if both } \mu \in I \text{ and } \nu \in I \\ \frac{1}{2}S_{\mu\nu} & \text{for other cases } (\mu \in I \text{ or } \nu \in I) \end{cases}$$

The $H$ matrix used in the secular equation (eq 4) becomes

$$H = H^0 + H^1 + V^I P^I$$

Similarly to the constrained DFT, eq 4 must now be solved self-consistently over the atomic charges and contains an unknown Lagrange multiplier $V^I$. To overcome some convergence problems, we have implemented three ways of solving this equation that can be used alternatively until one of them converges:

(i) The first one (similar to that of ref 63) consists of solving the secular equation with an inner loop and an outer loop. In the inner loop, the Hamiltonian is calculated with a fixed set of atomic charges, and the Lagrange multiplier $V^I$ is modified so that the MOs diagonalizing the Hamiltonian satisfy the charge localization constraint. The outer loop is the self-consistent loop over the atomic charges.

(ii) The second approach consists of inverting the two previous loops; i.e., the inner loop ensures the self-consistence over the Mulliken charges, and the external loop allows the determination of the Lagrange multiplier $V^I$.

(iii) The third approach is somewhat different and consists of three steps. First, a MOs guess is generated for the isolated fragments. The full set of MOs is orthonormalized with a Löwdin procedure. These MOs do not correspond to an energy minimum and do not satisfy the charge localization constraint. In the second step, the MOs evolve to change charge on fragment $I$ with the iterative procedure:

$$\phi_i^I(n+1) = \phi_i^I(n) + \alpha\left(p^I\phi_i^I(n) + \sum_j \phi_j^I(n)\Lambda_{ij}\right) \quad \forall i \quad (8)$$

where $n$ is the iteration step. The last term ensures the orthonormalization constraint. Transposing this equation in the atomic basis set gives the evolution of the MOs:

$$C^I(n+1) = C^I(n) + \alpha(S^{-1}P^IC^I(n) + XC^I(n)) \quad (9)$$

where $X = \alpha S^{-1}\Lambda$. At each step, the $\alpha$ coefficient is adapted to increase or decrease the charge on fragment $I$, and the $X$ matrix is calculated solving a second-order equation equivalent to the Rickaert algorithm[84] already implemented for SCC-DFTB Car−Parrinello molecular dynamics.[85] Once a solution satisfying the density constraint is achieved, the last step consists of relaxing the MOs to minimize the energy, under conservation of the charge localization and orthonormality constraints

$$\phi_i^I(n+1) = \phi_i^I(n) + \alpha\left(\frac{dE}{d\phi_i^{I*}} + \sum_j \phi_j^I\Lambda_{ij} + V^IP^I\phi_i^I\right) \quad (10)$$

giving the evolution of the coefficients

$$C^I(n+1) = C^I(n) + \alpha(S^{-1}HC^I(n) + XC^I(n) + V^IS^{-1}P^IC^I(n)) \quad (11)$$

This step requires both the calculation of $X$ and $V^I$. Starting from a given $V^I$ (the one in the previous step if $n > 1$), $C^I(n+1)$ is determined, calculating $X$ with the Rickaert algorithm. The charge carried by fragment $I$ with these new coefficients is calculated. If this charge is too large (respectively too small), $V^I$ is decreased (respectively increased). The process is repeated until the charge constraint is satisfied. Finally, the MOs converge to the charge-localized solution.

**2.3. The Configuration Interaction-Like Scheme.** The set of MOs $\{\phi_i^I\}$, obtained from a constrained SCC-DFTB calculation, is used to build the charge-localized configurations $\Phi^I$ as single Slater determinants. The coefficients $b_I$ of these configurations in the total wave function $\Psi$ (see eq 1) are obtained by solving the CI-like scheme:

$$\begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & .. & \mathbf{H}_{1n} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & .. & \mathbf{H}_{2n} \\ : & : & :: & : \\ \mathbf{H}_{n1} & . & .. & \mathbf{H}_{nn} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ : \\ b_n \end{pmatrix} = E\begin{pmatrix} \mathbf{S}_1 & \mathbf{S}_{12} & .. & \mathbf{S}_{1n} \\ \mathbf{S}_{21} & \mathbf{S}_2 & .. & \mathbf{S}_{2n} \\ : & : & :: & : \\ \mathbf{S}_{n1} & . & .. & \mathbf{S}_{nn} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ : \\ b_n \end{pmatrix} \quad (12)$$

where $\mathbf{S}_{IJ}$ is the two-configuration overlap $\langle\Phi^I|\Phi^J\rangle$ and $\mathbf{H}_{II}$ is the energy of the configuration $\Phi^I$ already calculated with the constrained SCC-DFTB. Following the approach of Wu et al.,[66,67] the coupling elements $\mathbf{H}_{IJ}$ are calculated by

$$\mathbf{H}_{IJ} = \frac{1}{2}(\mathbf{H}_{II} + \mathbf{H}_{JJ} + N^IV^I + N^JV^J)\mathbf{S}_{IJ} - \frac{1}{2}(V^I\langle\Phi^I|\hat{P}^I|\Phi^J\rangle + V^J\langle\Phi^I|\hat{P}^J|\Phi^J\rangle) \quad (13)$$

In the case of degenerate systems, one can also include more than one configuration to represent the charge localization on a given fragment, as will be shown in the applications of section 3. Solving eq 12 provides both the ground state of the system and some excited states generated via charge resonance. Although these excited states are also of interest, for instance, in spectroscopy, we focus in this work only on the ground state which corresponds to the lowest eigenvalue $E_g$.

**2.4. Analytical Gradients.** Derivatives of the energy with respect to atomic nuclear coordinates are required to perform molecular dynamics or geometry optimization. Their numerical calculation is possible by finite differences, but the number of energy calculations ($2 \times 3N_{\text{atoms}}$) turns out to be quite large, even for small systems. Thus, an analytical expression is of primary interest. In SCC-DFTB, it is convenient to use the derivatives of the matrix elements ($H^0$, $S$, $\Gamma$) which are known and tabulated. Differentiating the constrained SCC-DFTB energy with respect to the nuclear coordinate $\vec{R}_a$ of atom $a$ leads to the force expression

$$\vec{\nabla}_a E_g = \sum_{IJ} b_I b_J(\vec{\nabla}_a\mathbf{H}_{IJ} - E_g\vec{\nabla}_a\mathbf{S}_{IJ}) \quad (14)$$

We now present the calculation of derivatives for the diagonal and off-diagonal elements separately. The differentiation operator $\vec{\nabla}_a$ is replaced by the symbol $\partial_a$ to simplify the expressions.

*2.4.1. Derivative of the Diagonal Element.* The diagonal element $\mathbf{H}_{II}$ is the energy of the configuration $\Phi^I$. Differentiating the DFTB energy (eq 3 with the dispersion correction terms) and using the eigenvalue equation (eq 4), the molecular charge conservation (eq 7) and orthonormality constraints lead to the analytical expression (see also Wu and Van Voorhis[64]):

$$\partial_a\mathbf{H}_{II} = \partial_aE^{\text{rep}} + \sum_i n_i\sum_{\mu\nu} c_{i\nu}c_{i\mu}(\partial_aH^0_{\mu\nu} + V^I\partial_aP^I_{\mu\nu} + \left(\frac{H^1_{\mu\nu}}{S_{\mu\nu}} - \varepsilon_i\right)\partial_aS_{\mu\nu}) + q_a\sum_b \partial_a\Gamma_{ab}q_b + \partial_aE^{\text{disp}} \quad (15)$$

*2.4.2. Derivative of the off-Diagonal Elements.* The differentiation of the off-diagonal elements is obtained by differentiating eq 13, namely

$$\partial_a\mathbf{H}_{IJ;I\neq J} = \frac{1}{2}A_{IJ} + \frac{1}{2}A_{JI}$$

with

$$A_{IJ} = (\partial_a \mathbf{H}_{II} + N^I \partial_a V^I)\mathbf{S}_{IJ} + (\mathbf{H}_{II} + N^I V^I)\partial_a \mathbf{S}_{IJ} - \langle \Phi^I | P^I | \Phi^J \rangle \partial_a V^I - V^I \partial_a (\langle \Phi^I | P^I | \Phi^J \rangle) \quad (16)$$

We must now express the derivatives of the Lagrange multiplier $\partial_a V^I$ and those of the overlaps (real overlap and through the projectors) between $\Phi^I$ and $\Phi^J$. As there is no relationship between the MOs of the two configurations, there is no Hellman−Feynman type simplification for the derivatives of their overlaps. The analytical derivatives of the orbital coefficients and of the Lagrange multipliers must be explicitly calculated. The derivatives of the coefficients have already been expressed for DFT (see, for instance, ref 86) by solving the coupled perturbed equations. The expression only differs here through the term containing the constraint.

For a given configuration $\Phi^I$, the derivative of the coefficients of the orbitals $\{\phi_i^I\}$ can be related to the orbitals themselves through a $u$ matrix

$$\partial_a c_{i\mu}^I = \sum_k c_{k\mu}^I u_{ki} \quad (17)$$

The conservation of normalized MOs already imposes the form of the diagonal term of the $u$ matrix

$$u_{ii} = -\frac{1}{2} \sum_{\mu\nu} c_{i\mu}^I c_{i\nu}^I \partial_a S_{\mu\nu} \quad (18)$$

For the off-diagonal elements $u_{ij;\, i\neq j}$, differentiating eq 4 leads to

$$u_{ij} = \frac{\partial_a \mathscr{H}_{ij} - \varepsilon_j \partial_a \mathscr{S}_{ij}}{\varepsilon_j - \varepsilon_i} \quad (19)$$

where $\partial_a \mathscr{S}$ and $\partial_a \mathscr{H}$ are the derivatives of the SCC-DFTB overlap and Hamiltonian matrices expressed in the molecular orbital basis set

$$\begin{aligned} \partial_a \mathscr{S}_{ij} &= \sum_{\mu\nu} c_{i\mu}^I c_{j\nu}^I \partial_a S_{\mu\nu} \\ \partial_a \mathscr{H}_{ij} &= \sum_{\mu\nu} c_{i\mu}^I c_{j\nu}^I \partial_a H_{\mu\nu} \end{aligned} \quad (20)$$

In the constrained SCC-DFTB, the Hamiltonian matrix derivatives depend (i) on the derivatives of the matrices $H^0$, $S$, $\Gamma$, and $P$; (ii) on the derivatives of the coefficients; and (iii) on the derivatives of the Lagrange multipliers. These three contributions are now explicitly separated:

$$\partial_a \mathscr{H}_{ij} = \partial_a \mathscr{F}_{ij} + \sum_{kl} \mathscr{A}_{ij,kl} u_{lk} + \partial_a V^I \mathscr{P}_{ij} \quad (21)$$

where $\partial_a \mathscr{F}_{ij}$ contains the first contribution

$$\partial_a \mathscr{F}_{ij} = \sum_{\mu\nu} c_{i\mu}^I c_{j\nu}^I \partial_a F_{\mu\nu} \quad (22)$$

with

$$\partial_a F_{\mu\in\alpha,\nu\in\beta} = \partial_a H^0_{\mu\nu} + V^I \partial_a P_{\mu\nu} + \partial_a S_{\mu\nu} \frac{H^1_{\mu\nu}}{S_{\mu\nu}} + \frac{1}{2} S_{\mu\nu} \sum_\xi ((\partial_a \Gamma_{\alpha\xi} + \partial_a \Gamma_{\xi\beta}) q_\xi + \sum_i n_i \sum_l \sum_{\omega\in\xi} (\Gamma_{\alpha\xi} + \Gamma_{\xi\beta}) c_{i\omega} c_{il} \partial_a S_{\omega l})$$

The second term in eq 21 accounts for the Hamiltonian dependences on the orbital coefficients with

$$\mathscr{A}_{ij,kl} = \sum_{\mu\nu} c_{i\mu}^I c_{j\nu}^I \sum_\omega \frac{\partial H_{\mu\nu}}{\partial c_{k\omega}^I} c_{l\omega}^I \quad (23)$$

and

$$\frac{\partial H_{\mu\nu}}{\partial c_{k\omega}^I} = \frac{1}{2} n_k S_{\mu\nu} \sum_\xi \sum_{\lambda\in\xi} S_{\lambda\omega} (\Gamma_{\alpha\gamma} + \Gamma_{\beta\gamma} + \Gamma_{\alpha\xi} + \Gamma_{\beta\xi}) c_{k\lambda}^I \quad (24)$$

where $\mu \in \alpha$; $\nu \in \beta$; $\omega \in \gamma$. In the last term of eq 21, $\mathscr{P}_{ij}$ accounts for the Hamiltonian differentiation upon the Lagrange multiplier

$$\mathscr{P}_{ij} = \sum_{\mu\nu} c_{i\mu}^I c_{j\nu}^I P_{\mu\nu} \quad (25)$$

Compacting the $ij$ indices in a single $m$ index and the $kl$ indices in a single $n$ index, we now define

$$\begin{aligned} u_m &= u_{ij} \\ v_m &= \frac{\partial_a \mathscr{F}_{ij} - \varepsilon_j \partial_a \mathscr{S}_{ij}}{\varepsilon_j - \varepsilon_i} \\ B_{mn} &= \frac{A_{ij,kl}}{\varepsilon_j - \varepsilon_i} \\ w_m &= \frac{\mathscr{P}_{ij}}{\varepsilon_j - \varepsilon_i} \end{aligned}$$

and rewrite eq 19

$$\begin{aligned} u &= Bu + v + \partial_a V^I w \\ &= (1 - B)^{-1} v + \partial_a V^I (1 - B)^{-1} w \\ &= u^0 + \partial_a V^I u' \\ \text{with } u^0 &= (1 - B)^{-1} v \text{ and } u' = (1 - B)^{-1} w \end{aligned} \quad (26)$$

We now determine $\partial_a V^I$ using the fact that $N^I$ remains constant. Differentiating eq 7 leads to

$$\partial_a N^I = \sum_i n_i \sum_{\mu\nu} (c_{i\mu} c_{i\nu} \partial_a P_{\mu\nu} + 2 \partial_a c_{i\mu} c_{i\nu} P_{\mu\nu}) = 0 \quad (27)$$

which can be expressed with the $u$ matrix:

$$\partial_a N^I = \sum_i \sum_{\mu\nu} n_i c_{i\mu} c_{i\nu} \partial_a P_{\mu\nu} + 2 \sum_{ij} n_i u_{ji} \mathscr{P}_{ij} = 0 \quad (28)$$

Using the previous expression for $u$ leads to the following expression for the derivatives of the Lagrange multiplier:

$$\partial_a V^I = -\frac{\sum_i \sum_{\mu\nu} n_i c_{i\mu} c_{i\nu} \partial_a P_{\mu\nu} + 2\sum_{ij} n_i u^0_{ji} \mathscr{P}_{ij}}{2\sum_{ij} n_i u'_{ji} \mathscr{P}_{ij}} \quad (29)$$

$\partial_a \mathscr{H}$ can now be calculated from eq 21, as well as the $u$ matrix from eq 19, and finally the derivatives of the coefficients from eq 17. The $\mathbf{S}_{IJ}$ derivatives are computed from the MO coefficient derivatives and the derivatives of the AO overlap matrix. For sake of efficiency, the determinant expansions appearing in the calculation of the derivatives were calculated using the Sherman−Morrison formula.[87] A similar approach is applied to the derivatives of the projected overlap matrix $\langle \Phi^I | P^I | \Phi^J \rangle$.

Let us mention that, as a check, the analytical gradients have been compared to gradients obtained from finite difference calculations for a set of random geometries. The mean absolute value of forces was $1.4 \times 10^{-2}$ au with a root-mean-square of $1.8 \times 10^{-2}$ au, and the mean absolute error was $2.1 \times 10^{-5}$ au with a root-mean-square of $2.5 \times 10^{-5}$ au.

**2.5. Variant of the DFTB-VBCI: The HOMO Approximation.** We will consider the following approximation to the DFTB-VBCI approach: we assume that, in a molecular cluster, the MOs of the different charge localized configurations mostly differ through their Highest Occupied Molecular Orbital (HOMO). The overlaps and projected overlaps between two configurations can then be simplified as

$$\mathbf{S}_{IJ} = \langle \Phi^I | \Phi^J \rangle \simeq \langle \phi^I_{\mathrm{HOMO}} | \phi^J_{\mathrm{HOMO}} \rangle \quad (30)$$

$$\langle \Phi^I | P^I | \Phi^J \rangle \simeq N^I \langle \phi^I_{\mathrm{HOMO}} | \phi^J_{\mathrm{HOMO}} \rangle + \langle \phi^I_{\mathrm{HOMO}} | P^I | \phi^J_{\mathrm{HOMO}} \rangle$$

The off-diagonal CI matrix element becomes

$$\mathbf{H}_{IJ} \simeq \frac{1}{2}(\mathbf{H}_I + \mathbf{H}_J)\langle \phi^I_{\mathrm{HOMO}} | \phi^J_{\mathrm{HOMO}} \rangle - \frac{1}{2}(V^I \langle \phi^I_{\mathrm{HOMO}} | P^I | \phi^J_{\mathrm{HOMO}} \rangle + V^J \langle \phi^I_{\mathrm{HOMO}} | P^J | \phi^J_{\mathrm{HOMO}} \rangle) \quad (31)$$

The advantage of this approach is the ability to avoid any Slater determinant overlap calculation, and only the derivatives of the HOMO coefficients need to be calculated.

## 3. Applications

We will now apply the DFTB-VBCI method to two prototype cationic molecular clusters, namely, the benzene dimer and the water dimer. All calculations have been performed on a desktop computer (an Intel Xeon 2.8 GHz monoprocessor). Compared with DFTB, the DFTB+VBCI method is more time-consuming. For instance, a single point calculation (without gradients computation) for a water dimer performed over $6 \times 10^{-3}$ s at the DFTB level takes 0.1 s with the DFTB+VBCI. The single point calculation for the benzene dimer increases from 0.07 s at the DFTB level to 1.67 s at the DFTB+VBCI level. In an optimization procedure, the calculation of the gradient has a small effect on the water dimer (0.14 s per step) but a large effect on the benzene dimer (17 s. per step). Although the computational time is larger than for a simple DFTB calculation, which is the price to pay for treating the charge resonance effects correctly with this method, it remains much lower than high level *ab initio* methods, a typical optimization of about 100 steps for a benzene dimer taking half an hour.

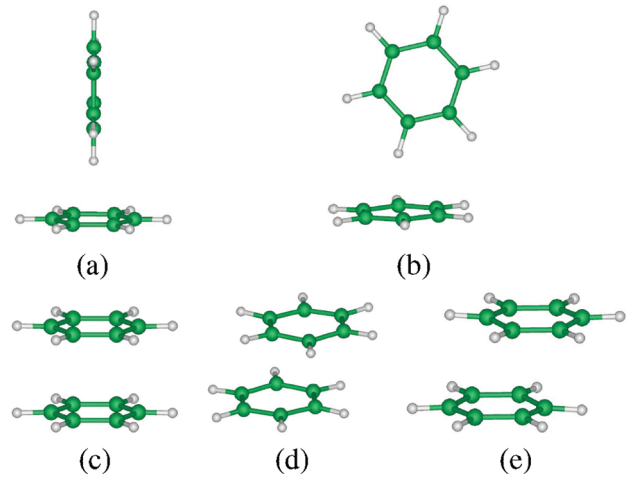



***Figure 1.*** Benzene dimer cations optimized at the DFTB-VBCI level. (a) T-shaped, (b) T_Csob, (c) sandwich stacked, (d) *x*-displaced, and (e) *y*-displaced isomers.

**3.1. The Cationic Benzene Dimer.** Several authors have investigated cationic benzene dimer clusters at high levels of theory, addressing the relative stability of characteristic isomers, namely, the sandwiches (stacked, parallel *x*- and *y*-displaced) and T-shaped configurations (see Figure 1 and Table 1). Let us cite for instance the work of Miyoshi et al.[88,89] which used a Complete Active Space Self-Consistent Field (CASSCF) followed by a Multi-Reference Coupled Pair Approximation (MRCPA) with a (73/7) basis set decontracted to (721/52) for C and Dunning's DZ set (31) for H. The sandwich parallel displaced isomers were found to be the most stable structures, with binding energies around 12.3 kcal $\mathrm{mol}^{-1}$, more stable than the T-shaped ones by 6.4 kcal $\mathrm{mol}^{-1}$. A similar study was performed by Pieniazeck et al.[90,91] with the Equation-Of-Motion Coupled-Cluster model with Single and Double substitutions for ionized systems (EOM-IP-CCSD/6-31+G*). This calculation yields the same isomer ordering as CASSCF-MRCPA and a similar energy difference between parallel displaced and T-shaped structures. The absolute binding energies are however much higher than for CASSCF-MRCPA (19.58 versus 12.3 kcal $\mathrm{mol}^{-1}$ for the *x*-displaced sandwich). In the following, the EOM-IP-CCSD results will be used as references to benchmark our model because (i) the structures have been fully optimized, whereas the CASSCF-MRCPA ones have only been optimized at the CASSCF level, and (ii) the binding energies are in good agreement with the experimental studies, providing values in the 15−20 kcal/mol range.[92−98]

The $D_{6h}$ symmetric stacking is another structure of interest, which is slightly less stable (about 1−2 kcal $\mathrm{mol}^{-1}$) than the two sandwich displaced isomers. We notice that DFT calculations[97,99,100] performed with the B3LYP functional give reasonable binding energies for the sandwich structure (17−19 kcal $\mathrm{mol}^{-1}$) but underestimate the energy difference between the two structures.

***Table 1.*** Binding Energies (kcal/mol) of the Cationic Benzene Dimer Obtained at Different Levels of Theory[a]

| | DFTB-VBCI | HOMO Approx. | SCC-DFTB | EOM-IP-CCSD | DFT | CASSCF + MRCPA |
|---|---|---|---|---|---|---|
| stacked sandwich | 17.70 | 17.91 | 29.53 | 18.34[b] | 18.2[d]−19.1[e] | |
| *x* displaced | 20.90 | 20.43 | 29.01 | 19.58[c] | | 12.3[g] |
| *y* displaced | 21.26 | 20.79 | 29.21 | 19.81[c] | 16.57[f] | 10.9[g] |
| T-shaped | 9.23 | 16.90 | 24.68 | 12.41[c] | 15.7[d] | |
| T_Csob | 9.19 | unstable | unstable | | | |

[a] The stacked sandwich structure correspond to constrained $D_{6h}$ optimization whereas the other isomers are fully optimized with the respective methods. [b] Pieniazek et al.[90] [c] Pieniazek et al.[91] [d] Ibrahim et al.[97] [e] Itagaki et al.[99] [f] Kryachko.[100] [g] Miyoshi et al.[89]

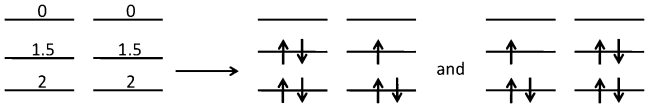



***Figure 2.*** Two electronic configurations (right) obtained from constrained SCC-DFTB calculation with noninteger occupation numbers (left).

Following Pieniazek et al.,[90] we call $\pi_g^a$ and $\pi_g^o$ the degenerate MOs in the neutral benzene molecule. In the ionized monomer, these two levels are degenerate at the neutral geometry but undergo Jahn−Teller distortion, leading to an acute angle configuration (ionization from the $\pi_g^a$ orbital) or an obtuse angle configuration (ionization from the $\pi_g^o$ orbital). We follow the electronic description of the Dimer Molecular Orbitals Linear Combination from the Fragment Molecular Orbitals (DMO-LCFMO[90]), to describe a benzene dimer, labeling the two fragments A and B. In this framework, the constrained state $A^+B$ can be obtained from removing one electron from either the $\pi_g^a$ or the $\pi_g^o$ orbitals of A. Consequently, we need two configurations to describe the constrained form $A^+B$. In the case of the symmetric $D_{6h}$ sandwich stacked dimer, these two $A^+B$ configurations are degenerate and are built as follows: we use the constrained SCC-DFTB to minimize the electronic energy with an occupation of 1.5 for the two highest occupied MOs (HOMO and HOMO−1) orbitals and 2 for the energetically lower lying orbitals. The two $A^+B$ configurations are then built from the obtained MOs as shown in Figure 2. The same procedure is applied to obtain two $AB^+$ configurations, and the CI matrix, which has to be diagonalized, is a $4 \times 4$ matrix.

In the other isomers (displaced sandwiches and T-shaped), the $\pi_g^a$ and $\pi_g^o$ orbitals of each fragment are no longer degenerate, and one could in principle calculate the energy of these configurations without using fractional occupation numbers. However, we could not obtain a self-consistent solution of the $A^+B$ state (respectively $AB^+$) with the constrained SCC-DFTB since the $\pi_g^a$ and $\pi_g^o$ orbitals on fragment A (respectively on B), although not degenerate, remain close in energy. We therefore decided to keep the procedure used for the $D_{6h}$ stacked sandwich isomer, filling the HOMO and HOMO−1 orbitals with 1.5 electrons. Although the filling of the MOs is fixed, these MOs relax anyway and are no longer degenerate in the final results due to the coupling with geometry relaxation. This fractional occupation of the HOMOs would also be usefull to describe the dissociation, keeping the same occupation of the orbitals (see ref 101). Finally, we mention that we use in the empirical dispersion term the parameters of Rapacioli et al.[79] already benchmarked for PAH clusters.

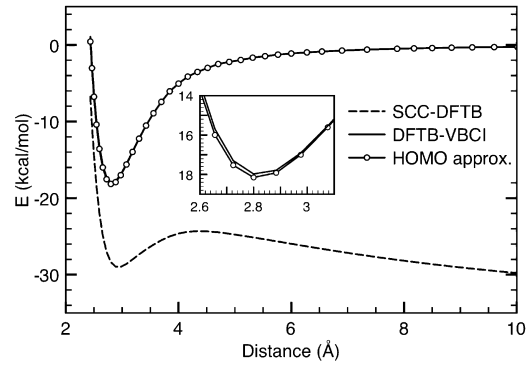



***Figure 3.*** Dissociation potential energy curves of the cationic benzene dimer in the stacked sandwich configuration calculated with SCC-DFTB and DFTB-VBCI approaches.

*3.1.1. The Stacked Sandwich Isomer.* We first discuss the results obtained for the stacked sandwich in the $D_{6h}$ geometry. Figure 3 represents the energy of the dimer corresponding to the dissociation along the *z* axis, orthogonal to the planes of the monomers. For this example, the fragments are frozen at the monomer neutral geometry. The zero energy reference corresponds to the sum of the separated fragments calculated independently, namely, $E(C_6H_6^+) + E(C_6H_6)$. The SCC-DFTB dissociation curve is reminiscent of the wrong dissociation curve of radical molecules like $H_2^+$ calculated with DFT (see for instance refs 35 and 36) and can be explained as follows. First, the SCC-DFTB energy does not converge to the sum of the energies of the fragments at the dissociation limit. At infinite distance, the charge is equally distributed over the two fragments. As the evolution of the self-interaction error with the number of electrons on a fragment is unfortunately not constant or linear, we have $2 \times E^{\text{SCC-DFTB}}(C_6H_6^{0.5}) \neq E^{\text{SCC-DFTB}}(C_6H_6^+) + E^{\text{SCC-DFTB}}(C_6H_6)$. At shorter distances, the energy increases, and a barrier is even observed before reaching the minimum, which is here a metastable minimum. The responsible repulsive contribution has a $1/R$ behavior and can be attributed to the artificial repulsion of two half-charged fragments, which is a different kind of self-interaction than the on site one. Finally, the minimum is much too low in energy as compared to the reference calculations (see Table 1). This overstabilization of delocalized states is a well-known effect of the self-interaction error.[102,103]

As can be seen from Figure 3, the DFTB-VBCI method does not present the wrong behavior pattern of the SCC-DFTB curve. At the dissociation limit, the energy converges to the sum of the energies of the fragments. In eq 12, the overlaps and coupling terms vanish, and the energies of the localized configurations are degenerate. These energies are

calculated with the electronic density corresponding to one charged and one neutral monomer and not that of two half-charged fragments. The Coulombic self-interaction $1/R$ repulsion also disappears with this approach, as well as the corresponding barrier. Finally, the binding energy for the stacked sandwich (17.70 kcal/mol) is significantly smaller than the SCC-DFTB one (29.53 kcal/mol) and yields a much better agreement with that of the EOM-IP-CCSD calculation at 18.34 kcal/mol. The interplane distance is 2.84 Å, which is smaller than the 3.3 Å reported at the EOM-IP-CCSD(T) level.[90]

The dissociation curve obtained by applying the HOMO approximation detailed in section 2.5 is also plotted in Figure 3. It is almost identical to the DFTB-VBCI curve with a binding energy of 17.91 kcal/mol vs 17.70 kcal/mol for the DFTB-VBCI.

*3.1.2. The T-Shaped and Displaced Sandwich Isomers.* The T-shaped and displaced sandwiches have been optimized without any geometrical constraint. The binding energies are reported in Table 1. The T-shaped isomer has a binding energy of 9.23 kcal/mol, which is slightly smaller than the EOM-IP-CCSD one. Another difference concerns the charge localization. With EOM-IP-CCSD, the charge is mostly localized on the stem fragment (88%), whereas its localization drops to 56% with the DFTB-VBCI. A possible explanation for this charge localization discrepancy could be related to some lack of stabilization by polarization. In DFTB-VBCI, the benzene $\pi$ system can be polarized in the direction parallel to the benzene ring. However, due to the reduced basis set used, the polarization of the $\pi$ system perpendicular to the benzene ring is underestimated. This lack of polarization could be at the origin of the destabilization of the configuration where the charge is carried by the stem fragment, leading to an oversharing of the charge and an underestimation of the binding energy. Another explanation could rely on the choice of the charge analysis method, which is a NBO analysis for the *ab initio* calculation and Mulliken analysis in the DFTB-VBCI. These two charge definitions are known to produce sometimes different charge distributions even for similar electronic densities. The distance between the centers of the two molecules is 4.52 Å, in good agreement with the EOM-IP-CCSD one (4.59 Å).

In the neutral benzene dimer, the most stable structures have often been reported to be "tilted" T-shaped (also called Cs over atom/bond) configurations.[10,79,104,105] In the cation, the corresponding structures have been reported to be transition states[100] (DFT-B3LYP level). Optimizing the Csob structure at the DFTB-VBCI level leads to a minimum (metastable, 11.7 kcal/mol above the global minimum). The energy difference with the T-shaped structure is smaller than 0.05 kcal/mol, certainly below the accuracy of the method.

At the SCC-DFTB level, the *x*- and *y*-displaced dimers are overstabilized as compared to reference calculations. Similarly to what is observed for the stacked sandwich dimer, the DFTB-VBCI approach gives considerably improved binding energies (20.90 and 21.26 kcal/mol) in very good agreement with those of EOM-IP-CCSD (19.58 and 19.81 kcal/mol). As already found for the $D_{6h}$ benzene case, the interplanar distance is shorter (2.71 and 2.78 Å) compared to EOM-IP-CCSD(T) results (3.08 and 3.18 Å). The sideward shiftings are 1.00 and 1.12 Å, compared to 1.07 (*x*-displaced) and 0.72 (*y*-displaced) Å at the EOM-IP-CCSD(T) level[91] (these shifts are 1.0 Å for both isomers when freezing Jahn−Teller relaxation[90]).

***Table 2.*** Binding Energies (kcal/mol) of Cationic Water Dimer Obtained at Different Levels of Theory

| | $[H_2O-H_2O]^+$ | $[H_3O-OH]^+$ ($C_s$) | $[H_3O-OH]^+$ ($C_1$) |
|---|---|---|---|
| DFTB-VBCI | 35.44 | 42.31 | |
| HOMO Approx. | 35.74 | 42.33 | |
| SCC-DFTB | 68.33 | 47.12 | |
| GGC | 53.73[a] | 48.66[a] | |
| BLYP | 58.4[b] | | 49.3[b] |
| B3LYP | 51.5[b] | | 49.8[b] |
| MPW1K | 42.9[b] | | 49.9[b] |
| BH&HLYP | 41.4[b] | | 49.9[b] |
| MP2 | 40.48[b]/43.5[c] | 50.9[c] | 46.47[b] |
| MP4 | 41.1[c] | 49.9[c] | |
| CCSD(T) | 39.53[b]/39.59[d] | 46.64[d] | 46.70[b]/46.68[d] |
| MCPF | 36.1[e] | 45.9[e] | 45.93[e] |

[a] Barnett and Landman.[112] [b] Lee and Kim.[111] [c] Gill and Radom.[106] [d] Cheng et al.[110] [e] Sodupe et al.[107] MCPF = SCF + electron correlation included with size extensive Modified-Coupled-Pair Functional.

The two structures are almost degenerate, with a slightly more stable *y*-displaced dimer. The energy difference is however probably much smaller than the expected accuracy of our approach. These are clearly cases in which quantum vibrational effects should be considered.

The global trend when relaxing the geometries compared to neutral dimers is to reduce intermolecular distances. For instance, the interplanar distances in displaced sandwich structures are 2.71/2.78 Å, smaller than those obtained for the neutral dimer (3.39 Å) at the DFTB level.[79] The sideward shifting is also reduced 1.0/1.13 Å versus 1.36 Å in the neutral dimer. The same trend is observed for the T-shaped isomer in which the distance between the molecular centers is reduced from 4.82 to 4.52 Å in the cationic dimer.

Applying the HOMO approximation to the DFTB-VBCI leads to very similar results. The most stable structures are the *x*- and *y*-parallel displaced ones with binding energies differing by less than 2.5% from those of DFTB-VBCI. The T-shaped structure is found to be less stable than the previous isomers, but its binding energy is overestimated as compared to reference calculations and DFTB-VBCI. This suggests that the differences between the charge localized configurations cannot be reduced to a change in the HOMO. We also notice that with this approximation the Csob does not correspond anymore to a minimum, the optimization leading to the T-shaped structure.

**3.2. The Cationic Water Dimer.** The potential energy surface of cationic water dimers has been investigated using high-level of theories[106−111] (see Table 2). The stable structures belong to two families (Figure 4). In the first one, the two water monomers are superimposed in an antisymmetric pattern, and the charge is equally distributed over the two units. The second one results from a proton transfer leading to two nonsymmetric units $[H_3O-OH]^+$, in which the charge is mostly localized on the $H_3O$ fragment. The structures found to be minima in ref 110

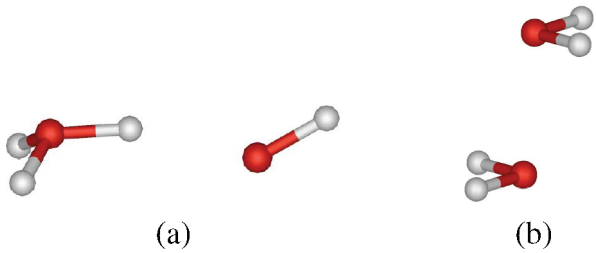



***Figure 4.*** Water dimer cations optimized at the DFTB-VBCI level. (a) $[H_3O-OH]^+$ isomer and (b) $[H_2O-H_2O]^+$ isomer.

have been taken as starts for optimization using the DFTB-VBCI method.

*3.2.1. The $[H_3O-OH]^+$ Isomer.* Table 2 compares the binding energies obtained at the SCC-DFTB level to those resulting from other calculations. For the $[H_3O-OH]^+$ isomer, most of the DFT functionals (except for the BH&H) give reasonable results as compared to CCSD(T) values. Similarly, the binding energy obtained at the SCC-DFTB level, without CI correction, is close to that of CCSD(T) (47.12 versus 44.6−46.7 kcal mol$^{-1}$).

In this system, DFTB-VBCI considers the interaction between the configurations where the positive charge is localized either on the $H_3O$ or on the OH fragments. At the dissociation limit, the constrained form ($H_3O^+-OH$) is obtained by fixing occupation numbers of 1.5 for the two degenerate HOMOs on the neutral OH fragment (in order to maintain the degeneracy of the OH $\pi_x$ and $\pi_y$ orbitals), the other occupied orbitals being doubly occupied. The second constrained form ($H_3O-OH^+$) is obtained with occupation numbers of 0.5 for the two degenerate HOMOs of the neutral $H_3O$ fragment and 0.5 the two degenerate HOMOs on the ionized OH fragment, the other orbitals being doubly occupied. In the complex, the degeneracies are lifted but, similarly to the benzene dimer, we decided to keep these fixed occupation numbers in order to prevent some convergency problems and to have a continuous description of the dissociation, which may be useful in future works.

The weights of the two configurations in the CI approach indicate that the charge is mostly localized (99.9%) on the $H_3O$ fragment. In CCSD(T) calculations, the charge is also strongly localized on this fragment but only by 88% from a restricted open-shell Hartree−Fock level with natural population analysis.[110]

The DFTB-VBCI minimum ($C_s$-trans) is different from the $C_1$ minimum obtained with CCSD(T). However, in CCSD(T), the $C_s$-trans isomer corresponds to a transition state 0.04 kcal/mol higher in energy than the global $C_1$ minimum[110] (0.1 kcal mol$^{-1}$ for the EOM-IP-CCSD[108] and 0.03 kcal/mol at the SCF+MCPF level[107]). Such a small energy difference is far beyond the expected accuracy of the DFTB-VBCI method. The binding energy of the $C_s$-trans isomer is close (42.31 kcal/mol) to that obtained with a simple SCC-DFTB calculation (47.12 kcal/mol). This is due to the fact that the charge is not significantly delocalized between the two fragments and that the SCC-DFTB calculation already attributes most of the charge to the $H_3O$ fragment. The artificial stabilization by the self-interaction error is therefore less crucial. This also explains why most of the DFT functionals give reasonable results for this structure. Concerning the geometry, the distance between the two oxygen atoms is 2.66 Å, close to the value of 2.5 Å usually found.[107,108,110,111] The hydrogen bonding is overestimated with 1.74 Å compared to values between 1.44 and 1.47 Å at a high level of calculations.[107,108,110,111]

*3.2.2. The $[H_2O-H_2O]^+$ Isomer.* It can be seen from Table 2 that, at the CCSD(T) level, the $[H_2O-H_2O]^+$ isomer is less stable by 7 kcal/mol than the $[H_3O-OH]^+$ isomer. At the DFT level, the binding energy strongly depends on the choice of the functional. For instance, the $[H_3O-OH]^+$ structure is more stable than $[H_2O-H_2O]^+$ with MPW1K, BH&H, and BH&LYP functionals, but it is the opposite with the BLYP, BPW91, HCTH407, and B3LYP functionals. At the SCC-DFTB level, the binding energy of the $[H_2O-H_2O]^+$ isomer is strongly overestimated (68 versus 39 kcal mol$^{-1}$ for CCSD(T)), making this isomer 23 kcal/mol more stable than the $[H_3O-OH]^+$ isomer. The DFTB-VBCI leads to a significant improvement, reducing the binding energy to 35.44 kcal/mol, a value close to CCSD(T) results (39 kcal/mol). In this isomer, the charge is equally distributed between the two equivalent fragments. The overstabilization observed at the SCC-DFTB level is attributed to the self-interaction error due to the strong delocalization and is corrected by the DFTB-VBCI approach. This is in line with the fact that self-interaction corrected functionals successfully predict this structure to be less stable by about 8 kcal mol$^{-1}$ than the proton transferred isomer.[108] The distance between the two oxygens is 2.05 Å, in agreement with values between 2.02 and 2.05 Å at higher levels of calculation.[107,108,110,111] We notice that our geometry corresponds to a $C_{2h}$ symmetry, whereas this optimized configuration is often reported in a $C_2$ geometry (see refs 108, 110, and 111). However, Cheng et al.[110] found that $C_2$ and $C_{2h}$ structures degenerate at the CCSD(T) level.

Finally, we notice (Table 2) that for both the $[H_3O-OH]^+$ and $[H_2O-H_2O]^+$ isomers, the binding energies obtained with the HOMO approximation are very close to that obtained with the full DFTB-VBCI method.

## 4. Conclusion

An extended method combining a VBCI-like scheme with SCC-DFTB has been developed. The method has been implemented together with its analytical gradients to enable complete optimization, including the intra- and intermolecular degrees of freedom.

We have benchmarked the DFTB-VBCI approach on the ionized dimers of benzene and water. It is shown for the benzene dimer cation that the self-interaction error is at the origin of the unphysical behavior of the SCC-DFTB dissociation energy curve. It is fully corrected with DFTB-VBCI, as detailed for the stacked sandwich. The binding energies obtained for different isomers with the DFTB-VBCI method agree well with those of high-level calculations as well as experimental data, while these energies are strongly overestimated with SCC-DFTB. We however notice that the main error for the DFTB-VBCI binding energy concerns the T-shaped structure, which is understabilized by 3 kcal/mol. This may be due to the use of point charges and a possible

mistreatment of the multipolar nature of the benzene $\pi$ system interacting with that of the charged stem benzene. Further improvement of the DFTB-VBCI could include such a multipolar description of the $\pi$ system as used, for instance, in accurate force field calculations[113] in order to account for this effect but at the price of a larger computational effort to derive the energy gradient.

The second benchmark system is the ionized water dimer. The two lowest energy isomers strongly differ by a proton transfer. The binding energy of the $[H_3O-OH]^+$ isomer calculated at the DFT level with several functionals is in good agreement with reference calculations. This is also the case with SCC-DFTB and DFTB-VBCI due the localization of the charge on the $H_3O$ fragment, reducing the multiconfigurational nature of the wave function and the self-interaction error in standard DFT-based calculations. On the contrary, in the $[H_2O-H_2O]^+$ isomer, the charge is equally carried by the two fragments, and the binding energies obtained at the DFT level strongly differ depending on the choice of the functional. With SCC-DFTB, this structure is overstabilized and becomes artificially the most stable one. This effect is corrected with the DFTB-VBCI approach, which gives a binding energy close to that of high-level calculations.

For the two benchmark systems, the binding energies are in quantitative agreement with those of higher levels of calculation. Concerning the geometries, some differences have been observed with those of high level calculations, the most critical one being the interplane distance in benzene sandwich structures. This could be due to the reduced basis used in DFTB leading to an underestimation of overlaps and consequently charge resonance stabilization at large distances. Neglecting the three body integrals in the DFTB could also play a role, which is difficult to estimate.

In this work, we have been concerned with the analytical derivation of the gradients, and optimizing the efficiency of the code will be a further step. One of the key computational difficulties is the double SCF involving both the charge and constraint. Efficiency could certainly be strongly improved by using extrapolation schemes of the Lagrangian parameter and atomic charges (Broyden[114] or Pulay[115] schemes) also transferring the SCF densities from one geometry to the next one. In the gradient computation, most of the time is spent in the calculation of the inverse of the *A* matrix, which could be calculated iteratively. Starting from the inverse of *A* calculated at the previous step would reduce the number of iterations. All of these improvements would of course not affect the accuracy of the method.

Beyond molecular clusters, the direct applicability of the method to the fragmentation of organometallic complexes might be less straightforward. The present scheme requires an *a priori* identification of the ligand metal partition, which may not be unique; then one possibility could be to use small-scale fragments, for instance, one per ligand. The present scheme has been applied to cationic systems. Dealing with the localization/delocalization process in anionic molecular clusters could be considered with a simillar scheme. However, the treatment of the molecular negative units is not very reliable since DFTB is expressed in a minimal valence basis, while the description of molecular anions even with DFT generally requires extended basis sets with diffuse functions and even sometimes very diffuse functions for describing dipole- and quadrupole-bound anions.

We now plan to perform global explorations of the potential energy surfaces through molecular dynamics or Monte Carlo sampling with the aim of studying ionized dimer dissociation. When the advantage of SCC-DFTB in terms of computational efficiency is taken, the DFTB-VBCI will allow for dealing with systems much larger than dimers. In our previous study, DFTB-VBCI has been used to characterize binding energies, ionization potentials, as well as charge localization in stacked coronene clusters with frozen intramolecular geometries and equal spacings between the units.[77] This preliminary work was however performed before the development of the analytical nuclear gradients, and it will be of interest to characterize the effects of intra- and intermolecular relaxation in these clusters. As for the hole delocalization, one could expect similar patterns as those of the rare gas clusters $He_n^+$, $Ne_n^+$, $Ar_n^+$, $Kr_n^+$, and $Xe_n^+$, for which the hole tends to delocalize on a few units (from 2 to 4, depending on the rare gas) and the other atoms tend to organize in crowns around a linear core.[116] We plan to investigate how the monomer internal degrees of freedom, the molecular extension, and the shape influence the size of the core unit and the general organization in molecular clusters with polyatomic monomers. Another perspective will be to study charge dynamics in such clusters, which is possible since the model also provides charge transfer excited states.

**Acknowledgment.** The authors would like to acknowledge the cluster research group GDR 2758 for its support and the supercomputing facility of Toulouse III University, CALMIP, for generous allocation of computer resources.

**Supporting Information Available:** Optimized geometries of cationic molecular dimers. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103–4.

(2) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401–4.

(3) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *123*, 104307–10.

(4) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 234114–12.

(5) Langreth, D. C.; Dion, M.; Rydberg, H.; Schroder, E.; Hyldgaard, P.; Lundqvist, B. I. *Int. J. Quantum Chem.* **2005**, *101*, 599–610.

(6) Chakarova-Kack, S. D.; Schroder, E.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. Lett.* **2006**, *96*, 146107–4.

(7) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. B* **2007**, *76*, 125112–11.

(8) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004–4.

(9) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908–6918.

(10) Gräfenstein, J.; Cremer, D. *J. Chem. Phys.* **2009**, *130*, 124105–16.

(11) Lewis, J. P.; Sankey, O. F. *Biophys. J.* **1995**, *69*, 1068–1076.

(12) Meijer, E. J.; Sprik, M. *J. Chem. Phys.* **1996**, *105*, 8684–8689.

(13) Gianturco, F. A.; Paesani, F.; Laranjeira, M. F.; Vassilenko, V.; Cunha, M. A. *J. Chem. Phys.* **1999**, *110*, 7832–7845.

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

(15) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

(16) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693–2699.

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

(18) Goursot, A.; Mineva, T.; Kevorkyants, R.; Talbi, D. *J. Chem. Theory Comput.* **2007**, *3*, 755–763.

(19) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.

(20) Pederson, M. R.; Heaton, R. A.; Lin, C. C. *J. Chem. Phys.* **1985**, *82*, 2688–2699.

(21) Krieger, J. B.; Li, Y. *Phys. Rev. A* **1989**, *39*, 6052–6055.

(22) Johnson, B. G.; Gonzales, C. A.; Gill, P. M. W.; Pople, J. A. *Chem. Phys. Lett.* **1994**, *221*, 100–108.

(23) Ruiz, E.; Salahub, D. R.; Vela, A. *J. Phys. Chem.* **1996**, *100*, 12265–12276.

(24) Goedecker, S.; Umrigar, C. J. *Phys. Rev. A* **1997**, *55*, 1765–1771.

(25) Baerends, E. J.; Gritsenko, O. V. *J. Phys. Chem. A* **1997**, *101*, 5383–5403.

(26) Csonka, G. I.; Johnson, B. G. *Theor. Chem. Acc.* **1998**, *99*, 158–165.

(27) Chermette, H.; Ciofini, I.; Mariotti, F.; Daul, C. *J. Chem. Phys.* **2001**, *115*, 11068–11079.

(28) Garza, J.; Vargas, R.; Nichols, J. A.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 639–651.

(29) Della Sala, F.; Gorling, A. *J. Chem. Phys.* **2001**, *115*, 5718–5732.

(30) Patchkovskii, S.; Ziegler, T. *J. Phys. Chem. A* **2002**, *106*, 1088–1099.

(31) Polo, V.; Gräfenstein, J.; Kraka, E.; Cremer, D. *Chem. Phys. Lett.* **2002**, *352*, 469–478.

(32) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771–1790.

(33) Polo, V.; Gräfenstein, J.; Kraka, E.; Cremer, D. *Theor. Chem. Acc.* **2003**, *109*, 22–35.

(34) Kummel, S.; Perdew, J. P. *Mol. Phys.* **2003**, *101*, 1363–1368.

(35) Gräfenstein, J.; Kraka, E.; Cremer, D. *J. Chem. Phys.* **2004**, *120*, 524–539.

(36) Gräfenstein, J.; Kraka, E.; Cremer, D. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1096–1112.

(37) Ciofini, I.; Adamo, C.; Chermette, H. *Chem. Phys.* **2005**, *309*, 67–76.

(38) Dinh, P. M.; Messud, J.; Reinhard, P. G.; Suraud, E. *Phys. Lett. A* **2008**, *372*, 5598–5602.

(39) Duch, W. *J. Mol. Struct. Theochem* **1991**, *234*, 27–49.

(40) Siegbahn, P. E. M. *The Configuration Interaction Method in Lecture Notes in Chemistry*; Roos, B. O., Eds.; Springer Verlag: New York, 1992; Volume *58*, pp 255−293.

(41) Roos, B. O. *The Multiconfigurational (MC) Self-Consistent Field (SCF) Theory in Lecture Notes in Chemistry*; Roos, B. O., Eds.; Springer Verlag: New York, 1992; Volume *58*, pp 177−254.

(42) Werner, H.-J. *Adv. Chem. Phys.* **1987**, *69*, 1.

(43) Bartlett, R. J. *Coupled-Cluster Theory: An Overview of Recent Developments*; Yarkony, D. R., Eds.; World Scientific: Singapore, 1995; pp 1047−1131.

(44) Helgaker, T.; Jorgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley & Sons: New York, 2000; pp 140−200.

(45) Savin, A. *Recent developments and applications of modern Density Functional Theory*; Seminario, J., Eds.; Elsevier: Amsterdam, 1996; pp 327−357.

(46) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.

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

(48) Goll, E.; Werner, H. J.; Stoll, H. *Chem. Phys.* **2008**, *346*, 257–265.

(49) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(50) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

(51) Schutz, M.; Werner, H.-J.; Lindh, R.; Manby, F. R. *J. Chem. Phys.* **2004**, *121*, 737–750.

(52) Heitler, W.; London, F. *Z. Phys.* **1927**, *44*, 455–472.

(53) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: New York, 1939; pp 183−220.

(54) Murrell, J.-N.; Kettle, S.; Tedder, J. *The Chemical Bond*; John Wiley & Sons: Chichester, U. K., 1985; pp 1−60.

(55) Shaik, S. S.; Hiberty, P. C. *A Chemist's Guide to Valence Bond Theory*; Wiley-Interscience: New Jersey, 2008; pp 1−290.

(56) Vragović, I.; Scholz, R. *Phys. Rev. B* **2003**, *68*, 155202–16.

(57) Bouvier, B.; Brenner, V.; Millié, P.; Soudan, J.-M. *J. Phys. Chem. A* **2002**, *106*, 10326–10341.

(58) Amarouche, M.; Durand, G.; Malrieu, J. P. *J. Chem. Phys.* **1988**, *88*, 1010–1018.

(59) Durand, G.; Spiegelman, F. *Theor. Chem. Acc.* **2006**, *116*, 549–558.

(60) Grigorov, M.; Spiegelman, F. *Surf. Rev. Lett.* **1996**, *3*, 211–215.

(61) Calvo, F.; Galindez, J.; Gadea, F. X. *Phys. Chem. Chem. Phys.* **2003**, *5*, 321–328.

(62) Calvo, F.; Bonhommeau, D.; Parneix, P. *Phys. Rev. Lett.* **2007**, *99*, 083401−4.

(63) Wu, Q.; Van Voorhis, T. *Phys. Rev. A* **2005**, *72*, 024502−4.

(64) Wu, Q.; Van Voorhis, T. *J. Phys. Chem. A* **2006**, *110*, 9212–9218.

(65) Wu, Q.; Van Voorhis, T. *J. Chem. Theory Comput.* **2006**, *2*, 765–774.

(66) Wu, Q.; Cheng, C.-L.; Van Voorhis, T. *J. Chem. Phys.* **2007**, *127*, 164119–9.

(67) Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *125*, 164105–9.

(68) Wu, Q.; Kaduk, B.; Van Voorhis, T. *J. Chem. Phys.* **2009**, *130*, 034109−7.

(69) Van Voorhis, T.; Kowalczyk, T.; Kaduk, B.; Wang, L.-P.; Cheng, C.-L.; Wu, Q. *Annu. Rev. Phys. Chem.* **2010**, *61*, 149–170.

(70) Roos, B. O. *Advances in Chemical Physics; Ab Initio Methods in Quantum Chemistry - II*; Lawley, K. P., Eds.; Wiley & Sons: Chichester, U. K., 1987; Volume *69*, pp 399−445.

(71) Werner, H. J. *Mol. Phys.* **1996**, *89*, 645.

(72) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev.B* **1995**, *51*, 12947–12957.

(73) Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185.

(74) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(75) Oliveira, A.; Seifert, G.; Heine, T.; duarte, H. *J. Braz. Chem. Soc.* **2009**, *20*, 1193–1205.

(76) Gräfenstein, J.; Cremer, D. *Theor. Chem. Acc.* **2009**, *123*, 171–182, 10.1007/s00214−009−0545−9.

(77) Rapacioli, M.; Spiegelman, F. *Eur. Phys. J. D* **2009**, *52*, 55–58.

(78) Zhechkov, L.; Heine, T.; Patchovskii, S.; Seifert, G.; Duarte, H. *J. Chem. Theory Comput.* **2005**, *1*, 841–847.

(79) Rapacioli, M.; Spiegelman, F.; Talbi, D.; Mineva, T.; Goursot, A.; Heine, T.; Seifert, G. *J. Chem. Phys.* **2009**, *130*, 244304–10.

(80) Heine, T.; Rapacioli, M.; Patchkovskii, S.; Frenzel, J.; Koster, A.; Calaminici, P.; Duarte, H. A.; Escalante, S.; Flores-Moreno, R.; Goursot, A.; Reveles, J.; Salahub, D.; Vela, A. deMon-Nano Experiment 2009. http://physics.jacobs-university.de/theine/research/deMon/ (accessed Nov 2010).

(81) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

(82) Lowdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365–375.

(83) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547–2553.

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

(85) Rapacioli, M.; Barthel, R.; Heine, T.; Seifert, G. *J. Chem. Phys.* **2007**, *126*, 124103–7.

(86) Wolff, S. *Int. J. Quantum Chem.* **2005**, *104*, 645–659.

(87) Hager, W. W. *SIAM Rev.* **1989**, *31*, 221–239.

(88) Miyoshi, E.; Ichikawa, T.; Sumi, T.; Sakai, Y.; Shida, N. *Chem. Phys. Lett.* **1997**, *275*, 404–408.

(89) Miyoshi, E.; Yamamoto, N.; Sekiya, M.; Tanaka, K. *Mol. Phys.* **2003**, *101*, 227–232.

(90) Pieniazek, P. A.; Krylov, A. I.; Bradforth, S. E. *J. Chem. Phys.* **2007**, *127*, 044317−16.

(91) Pieniazek, P. A.; Bradforth, S. E.; Krylov, A. I. *J. Chem. Phys.* **2008**, *129*, 074104−11.

(92) Field, F. H.; Hamlet, P.; Libby, W. F. *J. Am. Chem. Soc.* **1969**, *91*, 2839–2842.

(93) Grover, J. R.; Walters, E. A.; Hui, E. T. *J. Phys. Chem.* **1987**, *91*, 3233–3237.

(94) Krause, H.; Ernstberger, B.; Neusser, H. J. *Chem. Phys. Lett.* **1991**, *184*, 411–417.

(95) Meot-Ner, M.; Hamlet, P.; Hunter, E. P.; Field, F. H. *J. Am. Chem. Soc.* **1978**, *100*, 5466–5471.

(96) Hiraoka, K.; Fujimaki, S.; Aruga, K.; Yamabe, S. *J. Chem. Phys.* **1991**, *95*, 8413–8418.

(97) Ibrahim, Y.; Alsharaeh, E.; Rusyniak, M.; Watson, S.; Mautner, M. M. N.; El-Shall, M. S. *Chem. Phys. Lett.* **2003**, *380*, 21–28.

(98) Rusyniak, M.; Ibrahim, Y.; Alsharaeh, E.; Meot-Ner (Mautner), M.; El-Shall, M. *J. Phys. Chem. A* **2003**, *107*, 7656–7666.

(99) Itagaki, Y.; Benetis, N. P.; Kadam, R. M.; Lund, A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2683–2689.

(100) Kryachko, E. S. *Int. J. Quantum Chem.* **2007**, *107*, 2741–2755.

(101) Nesbet, R. K. *Proc. R. Soc. London* **1955**, *A230*, 312–321.

(102) Bally, T.; Sastry, G. N. *J. Phys. Chem. A* **1997**, *101*, 7923–7925.

(103) Lundberg, M.; Siegbahn, P. E. M. *J. Chem. Phys.* **2005**, *122*, 224103–9.

(104) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

(105) Lee, E. C.; Kim, D.; Jurecka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457.

(106) Gill, P. M. W.; Radom, L. *J. Am. Chem. Soc.* **1988**, *110*, 4931–4941.

(107) Sodupe, M.; Oliva, A.; Bertran, J. *J. Am. Chem. Soc.* **1994**, *116*, 8249–8258.

(108) Pieniazek, P. A.; VandeVondele, J.; Jungwirth, P.; Krylov, A. I.; Bradforth, S. E. *J. Phys. Chem. A* **2008**, *112*, 6159–6170.

(109) Periyasamy, G.; Levine, R. D.; Remacle, F. *Chem. Phys.* **2009**, *366*, 129–138.

(110) Cheng, Q.; Evangelista, F. A.; Simmonett, A. C.; Yamaguchi, Y.; Schaefer, H. F. *J. Phys. Chem. A* **2009**, *113*, 13779–13789.

(111) Lee, H. M.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 976–981.

(112) Barnett, R. N.; Landman, U. *J. Phys. Chem.* **1995**, *99*, 17305–17310.

(113) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353–10359.

(114) Broyden, C. *Math. Comput.* **1965**, *19*, 577–593.

(115) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556–560.

(116) Haberland, H.; von Issendorff, B.; Kolar, T.; Kornmeier, H.; Ludewigt, C.; Risch, A. *Phys. Rev. Lett.* **1991**, *67*, 3290–3293.

CT100412F

JCTC Journal of Chemical Theory and Computation

# Comparative Study on the Performance of Hybrid DFT Functionals in Highly Correlated Oxides: The Case of $CeO_2$ and $Ce_2O_3$

Jesús Graciani,[†] Antonio M. Márquez,[†] José J. Plata,[†] Yanaris Ortega,[†] Norge C. Hernández,[‡] Alessio Meyer,[§] Claudio M. Zicovich-Wilson,[||] and Javier Fdez. Sanz*,[†]

*Departamento de Química Física, Facultad de Química, Universidad de Sevilla, 41012 Sevilla, Spain; Departamento de Física Aplicada I, Universidad de Sevilla, 41011 Sevilla, Spain; Dipartimento IFM, Università di Torino, Via P. Giuria, 7, I-10125 Torino, Italy and Unitá INFM di Torino, Sezione F, via Giuria 5, I-10125 Torino, Italy; and Facultad de Ciencias, Universidad Autonoma del Estado de Morelos, Av. Universidad 1001, 62209 Cuernavaca, México*

Received August 3, 2010

**Abstract:** The outstanding catalytic properties of cerium oxides rely on the easy $Ce^{3+} \leftrightarrow Ce^{4+}$ redox conversion, which however constitutes a challenge in density functional based theoretical chemistry due to the strongly correlated nature of the 4*f* electrons present in the reduced materials. In this work, we report an analysis of the performance of five exchange-correlation functionals (HH, HHLYP, PBE0, B3LYP, and B1-WC) implemented in the CRYSTAL06 code to describe three properties of ceria: crystal structure, band gaps, and reaction energies of the $CeO_2 \rightarrow Ce_2O_3$ process. All five functionals give values for cell parameters that are in fairly good agreement with experiment, although the PBE0 hybrid functional is found to be the most accurate. Band gaps, 2*p*-4*f*-5*d* in the case of $CeO_2$ and 4*f*-5*d* in the case of $Ce_2O_3$, are found to be, in general, overestimated and drop off when the amount of Hartree−Fock exchange in the exchange-correlation functional decreases. In contrast, the reaction energies are found to be underestimated, and increase when the amount of HF exchange lowers. Overall, at its standard formulation, the B1-WC functional seems to be the best choice as it provides good band gaps and reaction energies, and very reasonable crystal parameters.

## 1. Introduction

Cerium oxides, either $CeO_2$ or nonstoichiometric $CeO_{2-x}$, hereafter referred to generically as ceria, have traditionally played the role of a support material in components of catalysts used in several chemical processes. Typical examples of industrial applications are the three-way catalysts in automotive catalytic converters, fluid-cracking catalysts in refineries, and ethylbenzene dehydrogenation catalysts used during the production of styrene.[1,2] Ceria is also an active component in low-temperature CO and VOC oxidation catalysts, wet-oxidation of organic pollutants in water, hydrocarbon-reforming and the water-gas-shift reaction. Initially, the promoting effect of ceria was attributed to the enhancement of the metal dispersion and the stabilization toward thermal sintering.[3,4] However, subsequent work has shown that ceria can act as a chemically active component as well, working as an oxygen reservoir able to deliver it in the presence of reductive gases and to incorporate it upon interaction with oxidizing gases.[5−7] Its ability to store, release, and transport oxygen ions indicates

* Corresponding author e-mail: sanz@us.es.

[†] Departamento de Química Física, Facultad de Química, Universidad de Sevilla.

[‡] Departamento de Física Aplicada I, Universidad de Sevilla.

[§] Dipartimento IFM, Università di Torino and Unitá INFM di Torino.

[||] Facultad de Ciencias, Universidad Autonoma del Estado de Morelos.

10.1021/ct100430q 
Published on Web 12/01/2010

that ceria is not just a mere spectator but it takes part in the catalytic reaction. For instance, in the case of oxidation reactions catalyzed by vanadia, the catalytic activity appears to be highly enhanced when supported on ceria as compared to more inert supports as silica and alumina.[8−11] Similar behavior is clearly seen in the case of the water−gas shift reaction, where experiments carried on $Rh/CeO_2$ and on pure $CeO_2$[12,13] reveal striking differences. Also, the very recent work of Park et al.[14] and of Rodriguez et al.[15,16] illustrates the importance of stabilizing $Ce^{3+}$ centers and the role of the ceria nanoparticles.

The outstanding properties of ceria, and, consequently, the broad use in heterogeneous catalysis are due to its facile $Ce^{3+} \leftrightarrow Ce^{4+}$ redox conversion,[17] however, the adequate description of the electronic configuration of $Ce^{3+}$ ions constitutes a challenge in density functional based theoretical chemistry due to the strongly correlated nature of the 4*f* electrons. Indeed, the 4*f* electrons in $Ce_2O_3$ are localized and the material behaves like a typical antiferromagnetic Mott−Hubbard insulator.[18] However, due to the well-known lack of cancellation of the Coulomb self-interaction, DFT approaches within the LDA or GGA frameworks predict metallic behavior.[19−26] To circumvent this problem within the DFT framework, the use of hybrid functionals, in particular the Heyd−Scuseria−Ernzerhof (HSE06) hybrid[27] has been recently reported for both fully reduced bulk $Ce_2O_3$,[23,28] and partially reduced $CeO_2$ (111) surfaces.[29] Alternatively, a much less demanding computational approach makes use of a Hubbard-like term to account for the strong *on-site* Coulomb interactions. Indeed, the choice of *U* is a delicate point as the physical idea behind the method is to improve the electron correlation description of an electron pair in a given orbital, and it is clear that the optimum *U* value for LDA and GGA can be different. Also, the *U* parameter has to be large enough to properly localize the 4*f* electron of $Ce^{3+}$, but without introducing undesired artifacts, such as overestimated band-gaps, and finally, as recently suggested by Castleton et al.,[25] the *U* value might be different for different properties under study. This latter aspect is not of minor importance as, for instance, the *U* value that better gives the lattice parameters must not necessarily provide the best energies or band gaps. Finally, one has to mention the possibility of using a *U* value determined in a self-consistent way: $U_{eff} = 5.30$ and 4.50 eV for LDA and GGA, respectively.[30,31] However, there is no guarantee that a self-consistent *U* will systematically improve calculated results. In this context, it is also worth mentioning that a recent work on lanthanide oxides using a many-body perturbation theory in the *GW*@LDA+*U* approach exhibits only a weak dependence on *U* in a physically meaningful range of *U* values.[32]

In spite of the empirical choice of the *U* parameter,[21,22] the DFT+*U* approach has been shown to be an effective, widely used, theoretical tool in the study of structure and reactivity of ceria surfaces. However, for accurate energies and properties, an approach without external semiempirical input appears to be preferable. For instance, let us consider the case of $CeO_2$ fluorite structure for which the experimental value is $a_0 \approx 5.41$ Å (5.406 Å[33] or 5.411 Å[34]). The LDA+U ($U_{eff} = 5.30$ eV) value is $a_0 = 5.40$ Å, in good agreement with the experiment, while GGA (PBE+U, $U_{eff} = 4.5$ eV)[23] moderately overestimates it: $a_0 = 5.49$ Å. This 1.3% error of the GGA represents a 4.5% increase in the equilibrium volume and it has been shown to be critical in the determination of the charge state of gold atoms deposited on $CeO_2$ (111) surfaces.[35−37] In its turn, a very accurate $a_0$ value results from hybrid DFT calculations: 5.39 and 5.40 Å from plane-wave calculations with the PBE0 and HSE hybrid functionals, respectively,[23] or 5.41 Å from calculations using a Gaussian-Type Orbitals (GTO) basis set and the HSE functional.[28]

The performance of both the DFT+*U* and the hybrid DFT approaches to describe the electronic properties of ceria has been analyzed in a series of papers. For instance, Hay et al.[28] compared the suitability of LDA, GGA and meta-GGA DFT functionals with HSE06 hybrid calculations using a GTO basis set. Furthermore, Da Silva et al.[23] compared PBE0 and HSE functionals using a plane wave basis set, and more recently, Kullgren et al.[38] have reported on the performance of B3LYP calculations. In the latter work, it was shown for instance that B3LYP performs slightly better than PBE0 for the electronic properties but slightly worse for the structural properties.

The work reported so far on the performance of the DFT functionals to describe the electronic properties of ceria makes it clear that hybrid functionals are better suited than DFT+*U* techniques to correctly render the structural and electronic properties of reduced ceria-based systems. Unfortunately, periodic hybrid DFT calculations face a number of computational problems that make them computationally demanding. Briefly, if we consider the plane-wave and GTO implementations, we find that energy calculations are reasonably fast when using GTO but geometry optimization becomes slow because the calculation of energy gradients in a GTO basis set becomes, generally speaking, the limiting step regardless the functional used. The choice of the basis set is also a key question especially for the 4*f* shell. In contrast, geometry optimizations are in general much more efficient when using a plane-wave basis set, but here the limiting step is the calculation of the energy with the hybrid functional due to the difficulty to estimate the nonlocal Fock exchange contribution. Finally, one must realize that hybrid DFT is sensitive to an additional factor because the amount of Fock exchange included in the potential is also an external input which largely affects the final description.[39,40]

Despite the recent efforts devoted to elucidate the properties of reduced ceria samples, and the ability of hybrid functionals to describe them, the body of literature about the subject still is scarce. In particular, most of the work has mainly been focused on structural and electronic aspects, while the energetic aspects, which are of major interest in chemistry, had not been in general considered. Moreover, a complete analysis of the dependence of the 4*f* band splitting, as well as the different band-gaps, cell parameters, and heats of formation on the amount of the exact exchange has not been yet reported. Indeed, as reported by Moreira et al. in their work on NiO,[40] the fraction of Fock exchange introduced in the hybrid functional does alter not only the

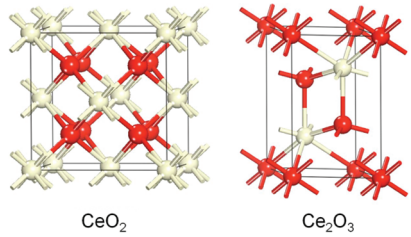



***Figure 1.*** Left: Fluorite type structure of $CeO_2$ (face-centered cubic, $Fm\bar{3}m$). Right: the sesquioxide A-type structure of $Ce_2O_3$ (hexagonal, $P\bar{3}m1$). Red and white balls indicate O and Ce atoms, respectively.

band-gaps but also the lattice constant, and the elastic constants and bulk modulus.

In view of the importance of this class of material, and the lack of information about the suitability of hybrid functionals to render a specific property, we have carried out in the present work a systematic analysis of the performance of five functionals commonly used in the literature, and that are implemented in the CRYSTAL06 code, namely the Perdew−Burke−Ernzerhof PBE0, the half and half HH, the modified half and half HHLYP, the widely used in computational chemistry B3LYP, and the recently proposed B1-WC functional. Using a purely ab initio periodic framework and treating oxidized and reduced ceria on an equal footing, we focus on the response of these functionals on three different sets of data: (i) structure: lattice parameters for $CeO_2$ and $Ce_2O_3$; (ii) band gaps: $2p$-$4f$-$5d$ in the case of $CeO_2$ and $4f$-$5d$ in the case of $Ce_2O_3$; and (iii) reaction energies involved in the $CeO_2/Ce_2O_3$ redox process. Moreover, bearing in mind the aforementioned sensitivity of the hybrid functionals to the fraction of exchange included, a systematic analysis of the behavior of the PBE0, B3LYP, and B1-WC functionals that incorporate different amounts of exact Fock exchange is also reported.

## 2. Computational Details

Two different structures were studied in this work, the $CeO_2$ fluorite crystal ($Fm\bar{3}m$) and the $Ce_2O_3$ A-type crystal ($P\bar{3}m1$). Their unit cells are shown in Figure 1. All of the calculations were performed using a developing version of the CRYSTAL06 code,[41] where the Fock (and Kohn−Sham, KS) equations[42] for the valence electron density are solved in a periodic framework. In this framework, the crystalline orbitals are represented as linear combinations of Bloch functions (BFs) and are evaluated over a regular three-dimensional mesh in the reciprocal space. Each BF is built from atom-centered atomic orbitals (AOs) that are contractions (linear combinations with constant coefficients) of Gaussian-type functions (GTFs), each GTF being the product of a Gaussian times a real solid spherical harmonic.

Five hybrid DFT functionals were used in this work: PBE0,[43] HH, HHLYP,[44] B3LYP,[45−47] and the recently proposed B1-WC functional.[48] Self-consistent-field (SCF) closed shell calculations were performed to obtain the ground electronic state in the case of $CeO_2$, while in the case of $Ce_2O_3$ spin-polarized calculations were performed in order to discriminate between the ferromagnetic and antiferromagnetic states of this oxide. In the latter, multiple solutions of the SCF take place depending on the accommodation of the unpaired electrons over the Ce $4f$ AOs. A recent implementation in the CRYSTAL program allows us to favor the convergence into a given symmetry adapted electronic configuration through a proper definition of the initial guess. In $Ce_2O_3$ calculations, the most stable configuration for the $4f$ electrons in Ce has been chosen. This is an antiferromagnetic state where both $\alpha$ and $\beta$ electrons occupy a mixing between $(2z^2 - 3x^2 - 3y^2)z$ and $(x^2 - 3y^2)x$ components of the $4f$ AOs of Ce.

The PBE0 is a combination of the GGA exchange-correlation functional PBE[49] ($E_{XC}^{PBE}$) and the exact Hartree−Fock (HF, $E_X^{HF}$) exchange following the expression:

$$E_{XC}^{PBE0} = E_{XC}^{PBE} + 1/4(E_X^{HF} - E_X^{PBE}) \quad (1)$$

The HH, HHLYP, B3LYP, and B1-WC follow the expression:

$$E_{XC} = (1\text{-A})(E_X^{LDA} + BE_X^{BECKE/WC}) + AE_X^{HF} + (1\text{-C})E_C^{VWN} + CE_C^{LYP/PBE} \quad (2)$$

where $E_X^{LDA}$ is the exchange contribution by using the Dirac−Slater functional[50] and $E_C^{VWN}$ is the correlation energy coming from the use of the Volsko−Wilk−Nusair parametrization of the Ceperley−Alder free electron gas correlation results.[51] In the case of HH, HHLYP, and B3LYP functional, $E_X^{BECKE/WC}$ stands for the Becke's exchange,[52] and $E_C^{LYP/PBE}$ represents the Lee−Yang−Parr correlation energy.[46] In the case of the B1-WC functional, $E_X^{BECKE/WC}$ stands for the Wu−Cohen[53] GGA exchange, and $E_C^{LYP/PBE}$ is the correlation energy contribution from the PBE.[49] Concerning the three weight parameters, $A = 0.2$, $B = 0.9$, and $C = 0.81$ for B3LYP. These parameters are set to $A = 0.5$ and $C = 1.0$ when we deal with the HH ($B = 0.0$) and HHLYP ($B = 1.0$) functional. In the case of using the B1-WC functional, $A = 0.16$ and $B = C = 1.0$.

Although calculations using the HSE06 functional are not performed in this work, we will also briefly outline it since is closely related to the PBE0 and largely used in the comparisons reported here. In the HSE functional, the spatial decay of the HF exchange interaction is accelerated by partitioning the Coulomb potential for exchange into short-range (SR) and long-range (LR) components:[27]

$$E_{XC}^{HSE} = aE_X^{HF,SR}(\omega) + (1\text{-a})E_X^{PBE,SR}(\omega) + E_X^{PBE,LR}(\omega) + E_C^{PBE} \quad (3)$$

where the mixing coefficient $a$ is set to 0.25, and the screening factor $\omega$ defines the separation range. This enables a substantial lowering of the computational cost for calculations in extended systems. Note that in the limit $\omega = 0$, HSE reduces to the hybrid functional PBE0, and when $\omega \rightarrow \infty$, *HSE becomes identical with PBE.*

Inner electrons of Ce atom were replaced by an effective core potential developed by the Stuttgart−Dresden group.[54] The Ce electrons explicitly treated were the $4s^2 4p^6 4d^{10} 5s^2 5p^6 4f^1 6s^2 5d^1$, with a (10sp7d8f)/[4sp2d3f] basis

***Table 1.*** Calculated and Experimental Lattice Parameters (in Å) for $CeO_2$ and $Ce_2O_3$

| | $CeO_2$ | | $Ce_2O_3$ | | | | |
|---|---|---|---|---|---|---|---|
| method | $a_0$ | error | $a_0$ | error | $c_0$ | error | refs |
| B3LYP | 5.47 | 0.06 | 3.89 | 0.00 | 6.17 | 0.11 | |
| HH | 5.34 | −0.07 | 3.83 | −0.06 | 5.93 | −0.13 | |
| HHLYP | 5.42 | 0.01 | 3.88 | −0.01 | 6.14 | 0.08 | |
| PBE0 | 5.40 | −0.01 | 3.86 | −0.03 | 6.04 | −0.02 | |
| B1-WC | 5.38 | −0.03 | 3.84 | −0.05 | 5.93 | −0.13 | |
| GGA(PBE)+U($U = 4.5$) | | | 3.87 | −0.02 | 5.93 | −0.13 | 24 |
| GGA(PW91)+U($U = 3.0$) | 5.48 | 0.07 | 3.92 | 0.03 | | | 22 |
| HSE | 5.41 | 0.00 | 3.87 | −0.02 | 6.06 | 0.00 | 23 |
| PBE0 | 5.39 | −0.02 | 3.87 | −0.02 | 6.07 | 0.01 | 23 |
| Experiment | 5.41 | | 3.89 | | 6.06 | | 33,34 |

set optimized to properly describe oxides where the metal features III and IV oxidation states. The corresponding exponents and coefficients can be found in ref 55. For O an all-electron basis set proposed in ref 56 for ionic crystals was adopted. The two most external *sp* and *d* exponents have been reoptimized for cerium oxide, their resulting values being 0.4798717, 0.1801227, and 0.2991812 $bohr^{-2}$, respectively.

Other technical parameters were set as follow. With the aim of obtaining an enough level of accuracy when evaluating the Coulomb and exchange series the five thresholds had the values of $10^{-8}$, $10^{-8}$, $10^{-8}$, $10^{-8}$, and $10^{-20}$. The Brillouin zone was sampled using a $6 \times 6 \times 6$ Monkhorst-Pack[57] grid, corresponding to 16 reciprocal space irreducible points at which the KS matrix was diagonalized. The SCF calculations were considered to be converged when the energy changes between the iterations were smaller than $10^{-8}$ hartree. The exchange-correlation contribution to the energy was evaluated by numerical integration over the cell volume.[58] Radial and angular points of the atomic grid were generated through Gauss−Legendre and Lebdev quadrature schemes. A grid pruning was adopted, as discussed in ref 58. In the present study, a (75, 974)p grid was used, such that it contains 75 radial points and a variable number of angular points, with a maximum of 974 on the Lebedev surface in the most accurate integration region. Full optimization (lattice constants and atomic positions) of $CeO_2$ and $Ce_2O_3$ were carried out using a convergence criterion of $3 \times 10^{-4}$ hartree/bohr for the root-mean-square values of forces and $1.2 \times 10^{-3}$ bohr in the root-mean-square values of atomic displacements. The Fermi level in the DOS plots is taken directly from CRYSTAL, and estimated in accordance with the zero-th level of the electrostatic energy in the multipolar Ewald-type expansion.[59]

## 3. Results and Discussion

**3.1. Crystal Structure.** By and large, DFT methods are known to predict fairly well the crystal structure of a wide variety of inorganic compounds. In general, the deviations of lattice parameters, both positive and negative, are in the range 2−3%,[60] hence it seems reasonable to adopt a value of 2.5% as accuracy criterion. Table 1 displays the computed lattice parameters for $CeO_2$ and $Ce_2O_3$ for each one of the functionals tested in this work. For comparison, some values chosen from the recent literature are also shown in this Table. In general, all computed values are found to correctly reproduce the experimental lattice parameters for both oxides, fulfilling the proposed accuracy criterion. The $a_0$ parameter for $CeO_2$ seems to be only modestly influenced by the exchange-correlation functional chosen. The largest errors correspond to the values computed with the B3LYP functional, the HH functional, or with the GGA+$U$ approach (∼1.3%, 0.06−0.07 Å), although they are well below the required accuracy criterion. Alternatively, the smallest errors are found for the HHLYP, HSE and PBE0 functionals (less than 0.4%).

The $a_0$ and $c_0$ lattice parameters for $Ce_2O_3$ show a similar behavior: the computed values are quite insensitive to the functional chosen, with all calculated values within the proposed error bar. The HH functional is, again, the one with the largest errors with respect to the experimental values, underestimating by 1.5% and 1.9% the lattice parameters. It is worth pointing out that, except for the HSE functional,[23] the percent errors on the computed $c_0$ lattice parameter are larger than the errors found for the calculated $a_0$ values. As in the case of the lattice parameter of $CeO_2$, the smallest average errors are found for the HHLYP, HSE, and PBE0 functionals (less than 0.6% on average). Finally, it is worth mentioning here that the computed values with the PBE0 functional are practically the same, no matter the kind of basis set used: plane waves[23] or localized atomic orbitals (this work).

**3.2. Electronic Structure.** *3.2.1. Electronic Structure of $CeO_2$.* In $CeO_2$, the valence and conduction band are mainly composed by O 2*p* and Ce 5*d* states, respectively, while the Ce 4*f* states lie within the gap. All valence Ce states, including the 4*f* states, are empty, and the system is a wide gap insulator (see Figure 2, top). All local, semi local, and hybrid functionals produce an insulating solution, in agreement with the above picture of the $CeO_2$ electronic structure. Besides this qualitative agreement, the theoretical description of the electronic structure of $CeO_2$ is quite sensitive to the approach used, as can be deduced from the different band gaps reported in Table 2. As expected, both LDA and PBE underestimate the main band gap (O 2*p*− Ce 5*d*).[23] However, it is interesting to note that all DFT+$U$ approaches reported in the literature also underestimate this band gap, and the results are not much sensitive to the specific value of the $U$ parameter.[22,23] This can be easily explained since the $U$ parameter acts only on the Ce 4*f* states, thus not modifying the relative positions of the valence and conduction bands, that have predominantly O 2*p* and Ce 5*d*

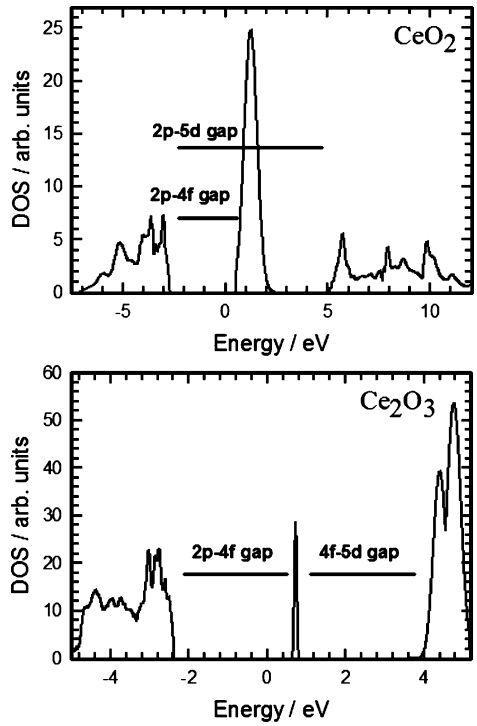



***Figure 2.*** Total density of states (DOS) for $CeO_2$ (top) and $Ce_2O_3$ (bottom).

***Table 2.*** Calculated and Experimental Band Gaps (in eV) for $CeO_2$

| method | O 2*p*−Ce 5*d* | O 2*p*−Ce 4*f* | Ce 4*f* −Ce 5*d* |
|---|---|---|---|
| B3LYP | 8.16 | 3.70 | 3.54 |
| HH | 10.64 | 7.50 | 1.91 |
| HHLYP | 10.75 | 7.18 | 2.51 |
| PBE0 | 8.52 | 4.30 | 3.10 |
| B1-WC | 7.48 | 3.18 | 3.09 |
| LDA[23] | 5.61 | 2.0 | 2.25 |
| PBE[23] | 5.64 | 2.0 | 2.5 |
| PBE0[23] | 7.93 | 4.5 | 2.25 |
| HSE[23] | 6.96 | 3.5 | 2.25 |
| HSE[28] | 7.0 | 3.3 | - |
| DFT+U[22,23] | ∼ 5 | - | - |
| Experiment[61,62,28] | ∼ 6−8 | 2.6−3.9 | |

character. Alternatively, all hybrid functionals are found to produce larger values of the O 2*p* - Ce 5*d* gap. Particularly, the HH and HHLYP functionals result in an overly large error (∼3−4 *eV*) with respect to the experimental value[61] for this band gap. This behavior might be ascribed to an excessive weight (50%) of the exact exchange in these two functionals. Among the different hybrid possibilities tested in this work, the B1-WC functional is the one that produces the smallest O 2*p*−Ce 5*d* gap, 7.48 eV, in close agreement with XPS and BIS experimental data,[62] which indicate a conduction band about 3 eV wide centered at about 7.5 eV. The smaller gap found for B1-WC agrees with the fact that it is the one incorporating the lowest HF exchange fraction. However, it is worth noting that although the HF fraction in B1-WC is lower than that of the screened HSE hybrid, the latter gives a gap even smaller (∼7 eV), in excellent agreement with the experimental data.

***Table 3.*** Calculated and Experimental Band Gaps (in eV) for $Ce_2O_3$

| method | O 2*p*−Ce 5*d* | O 2*p*−Ce 4*f* | Ce 4*f* −Ce 5*d* |
|---|---|---|---|
| B3LYP | 6.61 | 2.17 | 4.08 |
| HH | 9.4 | 1.83 | 7.19 |
| HHLYP | 9.78 | 1.13 | 8.25 |
| PBE0 | 7.08 | 2.34 | 4.54 |
| B1-WC | 5.94 | 3.00 | 2.78 |
| PBE0[23] | 6.75 | 3.25 | 3.50 |
| HSE[23] | 5.75 | 3.25 | 2.50 |
| experiment[18] | | | 2.40 |

The results found for the O 2*p*−Ce 4*f* gap closely follow the behavior previously discussed for the main band gap. The HH and HHLYP functionals produce band gaps that are too large, mainly because the excessive weight of the exact exchange pushes upward all virtual levels. Among the remaining results, the B1-WC hybrid functional (with the lowest HF fraction) produces the lowest band gap, 3.18 eV, again in close agreement with available experimental data,[62] and with the HSE values (3.3−3.5 eV), which also fall in the experimental range. The use of the PBE0 approach results in a band gap slightly larger that the experimental data, a behavior already reported and discussed.[23]

*3.2.2. Electronic Structure of $Ce_2O_3$.* In contrast to $CeO_2$, in $Ce_2O_3$ one electron per Ce atom populates the Ce band, resulting in a narrow 4*f* occupied band that develops in the O 2*p*−Ce 5*d* gap, some 2.4 eV below the conduction band[18] that is formed mainly by a mixing of Ce 5*d* and Ce 4*f* states (see Figure 2, bottom). Overall, the effect of the inclusion of the exact exchange in the hybrid functionals is similar to those found in $CeO_2$. The HH and HHLYP functionals result in too large band gaps, with all virtual levels too high in energy. With respect to the remaining hybrid functionals used in this paper, again the B1-WC produces the best result for the Ce 4*f*−Ce 5*d* gap (2.78 eV). This value agrees reasonably with the experimental value available (2.40 eV) and is slightly higher that the one estimated using the HSE functional (2.50 eV), which actually is the best to reproduce the experiment. In any case, except for the aforementioned cases of the HH and HHLYP functionals, the computed electronic structure of $Ce_2O_3$ is in semiquantitative agreement with the experimental information. Finally, if we compare the PBE0 band gaps obtained either with plane-wave or GTO basis sets noticeable differences might be seen, indicating that the electronic structure is more implementation dependent than the lattice parameters.

**3.3. Reaction Energies.** Given the active and crucial role played by $CeO_2$ and $Ce_2O_3$ oxides in many heterogeneous chemical reactions, generally traced back to their oxygen storage capacity, we have also investigated the performance of different hybrid functionals on the computation of some reaction energies involving cerium oxides. The suitability to predict the relevant thermodynamic properties has been investigated by computing the energetics of two reduction reactions involving $CeO_2$ and $Ce_2O_3$, namely:

$$2CeO_2(s) \rightarrow Ce_2O_3(s) + 1/2O_2(g) \qquad (3)$$

$$2\,CeO_2(s) + CO(g) \rightarrow Ce_2O_3(s) + CO_2(g) \quad (4)$$

The reaction enthalpies have been calculated as

$$\Delta H_1 = E(Ce_2O_3) + 1/2\,E(O_2) - 2\,E(CeO_2) \quad (5)$$

$$\Delta H_2 = E(Ce_2O_3) + E(CO_2) - 2E(CeO_2) - E(CO) \quad (6)$$

where $E$(X) represents the computed total energies of X = $CeO_2$ (solid), $Ce_2O_3$ (solid), $O_2$(gas), $CO_2$ (gas), and CO (gas) per formula unit. The experimental values have been obtained from the corresponding heats of formation of reactants and products.[63] The computed reduction energies, along with the experimental values are reported in Table 4. Zero-point vibrational energy contributions have not been included. A major problem that we encounter which makes it hard to extract any conclusions with respect to the reliability of the different functionals is that the experimental heats of formation of cerium oxides are not easy to measure. This problem is related to difficulties in the preparation of defect-free oxides in a well-defined oxidation state. Comparing the most recent reported data for reaction 3, shown in Table 4, with other values available in the recent literature (3.57 eV),[2] the uncertainty in the experimental value might be estimated to be ~0.5 eV. With this caution in mind we can, anyhow, comment on the values computed in this work with different hybrid functionals. The estimated reaction energies for reaction 3 show a dispersion (standard deviation) of 0.67 eV, with an average value of 3.92 eV, in close agreement with the last reported experimental value. Excluding the HHLYP value, that shows the larger absolute error with respect the experimental value, the average reaction energy increases to 4.12 eV (slightly larger than the experiment) and the dispersion is reduced to 0.50 eV, similar to the experimental error bar. Thus, regarding the computed reaction energies for reaction 3 we can say that all theoretical values fit within the experimental error bar, being the HH and PBE0 functional the ones that produce the data with the smaller deviation with respect to the currently accepted experimental reaction enthalpy. However, we can see once again a significant difference between the PBE0 estimated values obtained using plane-waves or GTO as basis set.

In the case of reaction 4, we can assume a similar error bar for the experimental value reported. Similar comments can be made with respect to the computed theoretical values. Excluding again the value obtained with the HHLYP functional (that predicts an exothermic reaction), the calculated average reaction energy will be 0.80 eV, in close agreement with the experimental reaction enthalpy of 1.02 eV. The standard deviation of the theoretical values is, in this case, only 0.36 eV, well within the experimental error bar. In this case, the HH and B1-WC functionals produce the theoretical values in better agreement with the experimental data.

***Table 4.*** Computed and Experimental Reaction Energies (in eV) for 2 $CeO_2 \rightarrow Ce_2O_3 + 1/2\ O_2$ and 2 $CeO_2$ + CO $\rightarrow Ce_2O_3 + CO_2$

| | 2 $CeO_2 \rightarrow Ce_2O_3 + 1/2\ O_2$ | | $CeO_2$ + CO $\rightarrow Ce_2O_3 + CO_2$ | |
|---|---|---|---|---|
| method | $\Delta H$ | error | $\Delta H$ | error |
| B3LYP | 3.52 | −0.47 | 0.44 | −0.58 |
| HH | 4.37 | 0.38 | 0.93 | −0.09 |
| HHLYP | 2.88 | −1.11 | −0.16 | −1.18 |
| PBE0 | 3.66 | −0.33 | 0.40 | −0.62 |
| B1-WC | 4.45 | 0.46 | 1.05 | 0.03 |
| PBE0[23] | 3.14 | −0.85 | | |
| HSE[23] | 3.16 | −0.83 | | |
| LDA+U[23] | 3.04 | −0.95 | | |
| PBE+U[23] | 2.29 | −1.70 | | |
| experiment[63] | 3.99 | | 1.02 | |

In summary, we can state that, excluding the HHLYP functional, all tested functionals produce values for the reaction energy that are within the experimental error bar of 0.5 eV.

**3.4. Effect of the Fock Exchange.** In addition, to evaluate the performance of different hybrid functionals in the description of the geometric and electronic structure of $CeO_2$ and $Ce_2O_3$, and in selected reaction energies involving ceria, we have also investigated to what extent the amount of HF exchange affects the three properties that we are looking at in this work: the cell parameters, the band gaps, and the reaction energies involved in the $Ce^{3+} \leftrightarrow Ce^{4+}$ redox process. Taking into account the results we have obtained so far, we have limited this analysis to the B3LYP, PBE0, and B1-WC functionals.

*3.4.1. Cell Parameters.* Figure 3 shows the influence of the amount of exact exchange in the computed values of the lattice parameters of $CeO_2$ and $Ce_2O_3$. Starting with the $a_0$ parameter of $CeO_2$, the computed lattice parameter decreases in all cases on increasing the percentage of HF exchange. This first result is in contrast with that reported on NiO, where the lattice constant was found to increase when the amount of exact exchange was raised.[40] For the PBE0 functional, the value computed with the standard amount of exact exchange (25%) is already very close to the experimental value, while for the B3LYP functional, the experimental lattice parameter is only reached at ~55% of exact exchange and for the B1-WC functional the most accurate value is obtained at 0% of exact exchange.

With respect to the lattice parameters of $Ce_2O_3$, we find that the $a_0$ parameter is quite insensitive to the amount of exact exchange in the three functionals tested. Only for the B3LYP functional, the experimental value of $a_0$ is reached (for 10−20% of HF exchange), while for PBE0 and B1-WC the computed value is always below the experimental one. The $a_0$ lattice parameter remains almost invariant with the B1-WC functional: the absolute change is less that 0.01 Å in the tested range (10−50% of HF exchange). The value of $c_0$ is more sensitive than that of $a_0$ to the fraction of HF exchange included in the hybrid functional. For the B3LYP and PBE0 functionals, the computed value decreases on increasing the participation of HF exchange, while for the B1-WC functional $c_0$ increases slightly. In the case of the PBE0 functional, the value closest to the experimental data is reached at about 10% of exact exchange, even though, with the standard value of HF exchange the error is only −0.35%, which is quite accurate and keeps the advantages of using a standard definition of the functional. In the case of the B3LYP functional, the experimental value of $c_0$ is

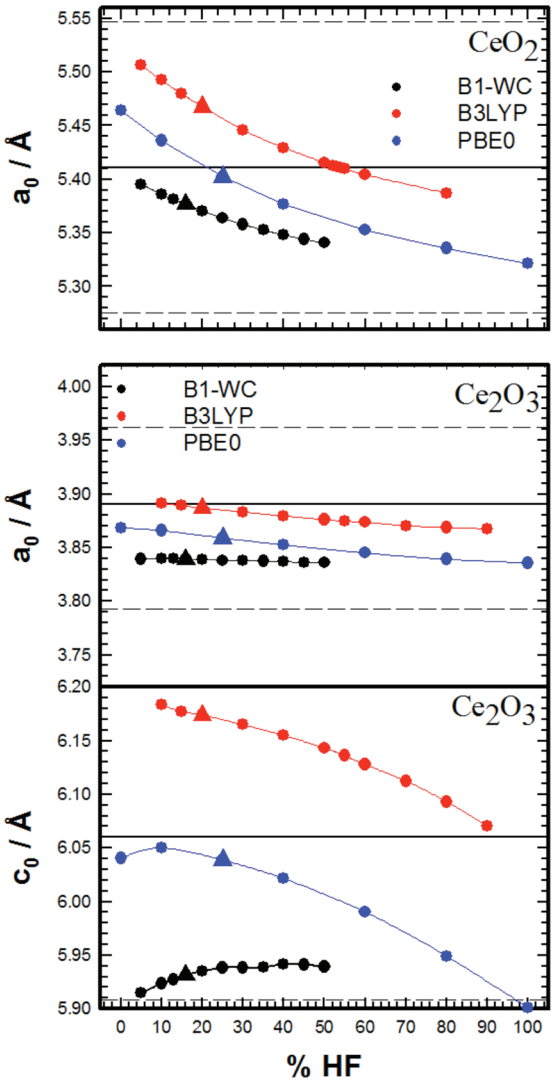



***Figure 3.*** Dependency of the computed lattice parameters for $CeO_2$ (top) and $Ce_2O_3$ ($a_0$, middle and $c_0$, bottom) on the amount of exact exchange in the B3LYP, PBE0, and B1-WC functionals. The triangles denote the standard HF percent for each functional. The horizontal lines show the experimental value (full) and the acceptable error bars (dashed).

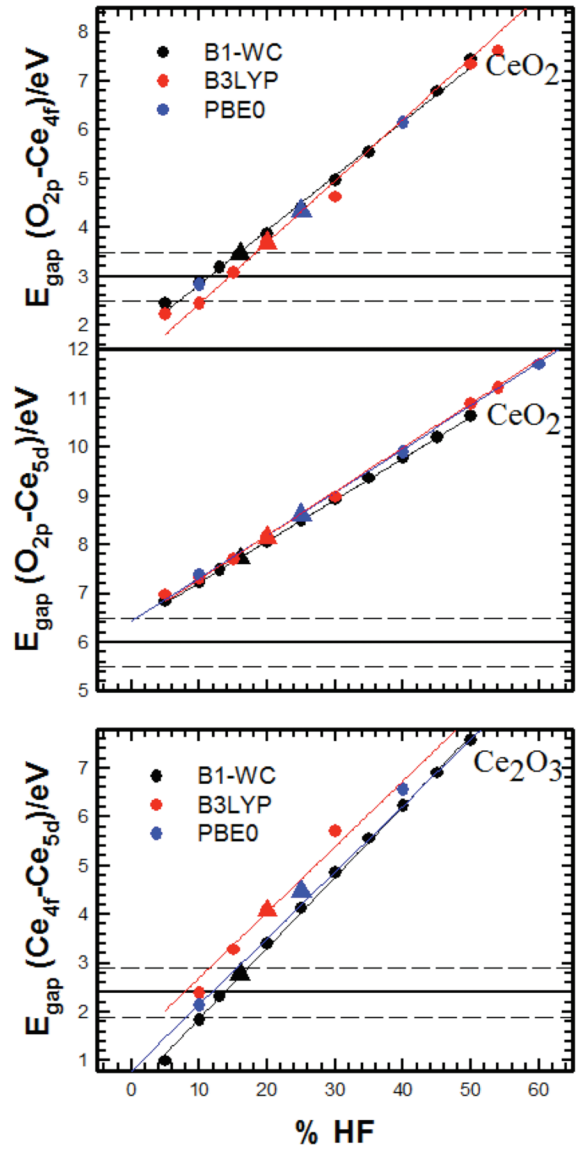



***Figure 4.*** Dependency of the computed band gaps for $CeO_2$ (top and middle) and $Ce_2O_3$ (bottom) on the percentage of exact exchange in the B3LYP, PBE0, and B1-WC functionals. Refer to Figure 3 for labeling.

not reached for any amount of exact exchange, being only approximated when the amount of HF exchange if increased up to 90%. The variation of $c_0$ with the percentage of HF exchange in the B1-WC functional is (as it did happened with $a_0$) quite small. The computed lattice parameter increases only slightly in the tested range of exact exchange, the total increment being less than 0.05 Å.

In summary, the analysis of the variations in the cell structure suggests that overall the PBE0 at its original formulation is the more appropriate choice to simultaneously render the three parameters. Although in the case of the sesquioxide the B3LYP answer for $a_0$ is very good, it needs to incorporate a large amount of Fock exchange to get closer to the experimental $c_0$ value. In the general comparison the B1-WC functional seems to perform satisfactorily with small variations.

*3.4.2. Band Gaps.* The evolution of the computed band gaps with the amount of Fock exchange included in the functional shows, for the three functionals tested, a marked linear behavior (see Figure 4). In all cases, and in agreement with the well-known trend,[40,64] the computed band gaps increase with the amount of exact exchange, the increment being practically linear, showing similar slopes, with many of the fitting lines overlapping. A direct consequence of this behavior is that similar values of the band gaps are computed for similar contributions of the Fock exchange, regardless of the functional utilized. The O 2*p*−Ce 4*f* experimental band gap of $CeO_2$ is, thus, most approximated in the 10−15% range of exact exchange. On the contrary, the computed O 2*p*−Ce 5*d* band gap of $CeO_2$ is always larger than the experimental value of 6.0 eV,[61] but it approaches the ∼7.0−7.5 eV value[62] obtained from XPS and BIS data, for about 10% Fock exchange. In the case of $Ce_2O_3$, the experimental value is, again, more closely approached in the 10−15% range of exact exchange and the computed value increases linearly with increasing contribution of the HF exchange in the functional. It is particularly striking again

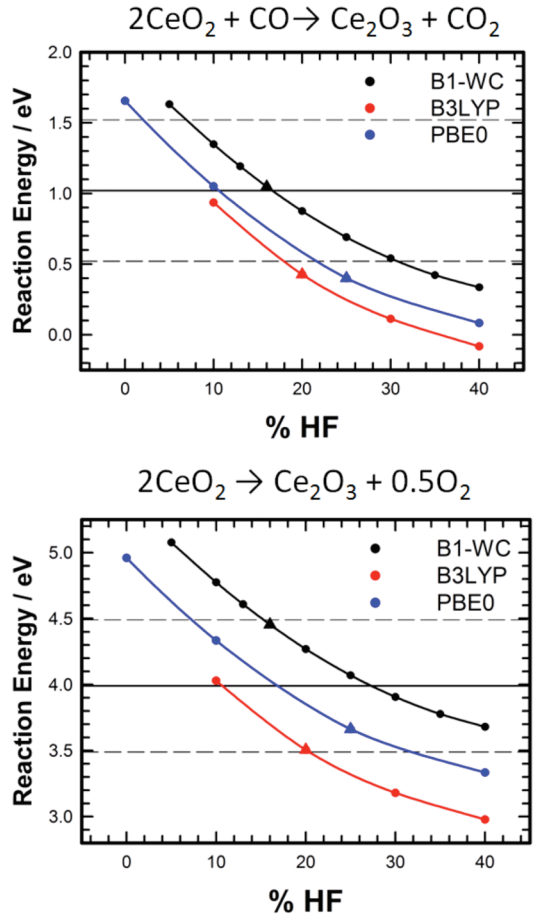



***Figure 5.*** Dependency of the computed reaction energies on the percentage of exact exchange in the B3LYP, PBE0, and B1-WC functionals. Refer to Figure 3 for labeling.

that, regardless of the functional used, all experimental band gaps are nearly approximated in a similar range of HF exchange, around 10−15%. This can be interpreted in the sense that the different band gaps (the relative position of the bands) are mostly dependent on the exchange functional and basically independent of the correlation functional utilized.

Summarizing again, when we analyze the plots for the three band gaps reported in Figure 4, the B1-WC functional is the one that gives a better result. Only this functional at its original formulation is able to reproduce band gaps within the ±0.5 eV error bar, even though the O 2*p*−Ce 5*d* gap still is overestimated.

*3.4.3. Energetics.* The dependence on the computed reduction energies for reactions 3 and 4 on the percentage of HF exchange included in the functional is shown in Figure 5. In both cases, the estimated reaction energies are decreased on increasing the contribution of exact exchange and show a similar dependency as demonstrated by the curve fitted to the computed reaction energies. It is always possible to adjust the contribution of HF exchange to reproduce the experimental reaction energy. For the reduction of $CeO_2$ with CO, reaction 4, amounts of HF exchange of 10−15% seem to give the best agreement with the experimental reaction energy. For reaction 3, however, the dispersion of the computed reaction energy is larger and, as result, the percentage of HF exchange required to for the experimental value also spawns a larger range: ∼10% for B3LYP, ∼18% for PBE0, and ∼28% for the B1-WC functional. At its original formulation, the B1-WC functional is once again the best well-behaved functional as it is very accurate for reaction 4 and within the bar error in the case of reaction 3.

***Table 5.*** Summary of the Effect of the Fock Exchange Contribution on the Structure, Band Gaps and Reaction Energies[a]

| property | error limit | B3LYP(20) | PBE0(25) | B1WC(16) |
|---|---|---|---|---|
| structure (cell parameters) | 2.5% | [0−100] | [0−100] | [0−100] |
| band-gaps $CeO_2$ | 0.5 eV | [10−18] | [8−16] | [8−16] |
| band-gaps $Ce_2O_3$ | 0.5 eV | [4−12] | [8−16] | [10−16] |
| energy for reaction 3 | 0.5 eV | [8−20] | [8−32] | [16−50] |
| energy for reaction 4 | 0.5 eV | [2−18] | [2−22] | [7−30] |
| the whole set | | [10−12] | [8−16] | 16 |

[a] Values that fulfill the accuracy criteria are in brackets, and the standard fractions between parentheses.

A complete view of the final ranges of possible HF fractions for which the different calculated values fall within the error bars might be obtained inspecting Table 5. As can be seen, the fraction for a given functional giving results within the error bars for lattice constants, band gaps, and energetics may differ, however, the three hybrids with the same nominal fraction of roughly 10−16% give results in fairly good agreement with the experimental data.

## 4. Conclusions

In this work, we report an analysis of the performance of five exchange-correlation functionals implemented in the CRYSTAL06 code to describe three properties of ceria that include crystal structure, band gaps, and reaction energies involved in the $Ce^{3+} \leftrightarrow Ce^{4+}$ redox process. Concerning the cell parameters, all five functionals give values that are within the 2.5% error, usually found for a vast majority of inorganic crystals, although the PBE0 hybrid functional is found to be the most accurate, giving parameters also very close to those estimated using the HSE screened functional. In general, when the fraction of HF exchange increases, a moderate lowering in the cell parameters is observed. Things change when we look at the band gaps of both cerium dioxide and sesquioxide. First the HH and HHLYP functionals lead to band gaps too large as they incorporate too much HF exchange that pushes the empty states too high. It is shown that for any functional used, the gaps are overestimated, and the agreement improves lowering the amount of the HF exchange. In this case, the overall best answer is provided by the B1-WC functional, which actually, among the functionals here considered, incorporates the lowest amount of HF exchange in its original formulation. The suitability to render the reaction energies normally involved in the rich ceria chemistry has been evaluated estimating the energetics associated to the $CeO_2 \rightarrow Ce_2O_3$ reduction process. For the two reactions considered, the reaction energies are in general underestimated, and lower when the amount of HF exchange increases, which is in contrast with the gaps behavior. Overall, the B1-WC functional is once again the most well-behaved functional to reproduce the correct energetics.

In summary, the present work shows that as far as the structure is concerned, any of the functionals that we have considered, in the original formulation, are accurate enough, giving parameters within the usual error bar. In general, the cell parameters are found to depend only moderately on the HF exchange fraction. However, caution should be taken in the case that the structure might favor a given state or property, in which case the PBE0 functional should be the choice. In the case of band gaps and reaction energies, a stronger dependency on the amount of HF exchange is observed. Its lowering improves band gaps and reaction energies with both PBE0 and B3LYP functional. Otherwise, at its standard formulation, the B1-WC functional (the one with the smallest fraction of HF exchange), seems to be the best choice as it provides good band gaps and reaction energies, and very reasonable crystal parameters.

**Acknowledgment.** This work has been supported by the Spanish Ministry of Science and Innovation, MICINN (Grant Nos. MAT2008-04918, CSD2008-0023), and the Junta de Andalucía (P08-FQM-3661). Computational time on the Barcelona Supercomputing Center/Centro Nacional de Supercomputación is gratefully acknowledged.

## References

(1) Trovarelli, A. *Catal. Rev. Sci. Eng.* **1996**, *38*, 439.

(2) Trovarelli, A. *Catalysis by Ceria and Related Materials; Catalytic Science Series*; Imperial College Press: London, 2002; Vol. 2.

(3) Dictor, R.; Roberts, S. *J. Phys. Chem.* **1989**, *93*, 5846.

(4) Su, E. C.; Rothschild, W. G. *J. Catal.* **1986**, *99*, 506.

(5) Yao, H. C.; Yu Yao, Y. F. *J. Catal.* **1984**, *86*, 254.

(6) Engler, B.; Koberstein, E.; Schubert, P. *Appl. Catal.* **1989**, *48*, 71.

(7) Miki, T.; Ogawa, T.; Haneda, M.; Kakuta, N.; Ueno, A.; Tateishi, S.; Matsuura, S.; Sato, M. *J. Phys. Chem.* **1990**, *94*, 6464.

(8) Daniell, W.; Ponchel, A.; Kuba, S.; Anderle, F.; Weingand, T.; Gregory, D. H.; Knozinger, H. *Top. Catal.* **2002**, *20* (1−4), 65–74.

(9) Wachs, I. E. *Catal. Today* **2005**, *100*, 79–94.

(10) Dinse, A.; Frank, B.; Hess, C.; Habel, D.; Schomacker, R. *J. Mol. Catal. A* **2008**, *289*, 28–37.

(11) Ganduglia-Pirovano, M. V.; Popa, C.; Sauer, J.; Abbott, H.; Uhl, A.; Baron, M.; Stacchiola, D.; Bondarchuk, O.; Shaikhutdinov, S.; Freund, H.-J. *J. Am. Chem. Soc.* **2010**, *132*, 2345.

(12) Shido, T.; Iwasawa, Y. *J. Catal.* **1992**, *136*, 493.

(13) Shido, T.; Iwasawa, Y. *J. Catal.* **1993**, *141*, 71.

(14) Park, J. B.; Graciani, J.; Evans, J.; Stacchiola, D.; Ma, S. G.; Liu, P.; Nambu, A.; Sanz, J. F.; Hrbek, J.; Rodriguez, J. A. *Proc. Natl. Acad. Sci.* **2009**, *106*, 4975.

(15) Rodriguez, J. A.; Graciani, J.; Evans, J.; Park, J. B.; Yang, F.; Stacchiola, D.; Senanayake, S. D.; Ma, S.; Perez, M.; Liu, P.; Sanz, J. F.; Hrbek, J. *Angew. Chem., Int. Ed.* **2009**, *48*, 8047.

(16) Park, J. B.; Graciani, J.; Evans, J.; Stacchiola, D.; Senanayake, S. D.; Barrio, L.; Liu, P.; Sanz, J. F.; Hrbek, J.; Rodriguez, J. *J. Am. Chem. Soc.* **2010**, *132*, 356.

(17) Trovarelli, A.; de Leitenburg, C.; Boaro, M.; Dolcetti, G. *Catal. Today* **1999**, *50*, 353.

(18) Prokofiev, A.; Shelykh, A.; Melekh, B. *J. Alloys Compd.* **1996**, *242*, 41.

(19) Yang, Z.; Woo, T. K.; Baudin, M.; Hermansson, K. *J. Chem. Phys.* **2004**, *120*, 7741.

(20) Nolan, M.; Grigoleit, S.; Sayle, D. C.; Parker, S. C.; Watson, G. W. *Surf. Sci.* **2005**, *576*, 217.

(21) Andersson, D. A.; Simak, S. I.; Johansson, B.; Abrikosov, I. A.; Skorodumova, N. V. *Phys. Rev. B* **2007**, *75*, 035109.

(22) Loschen, C.; Carrasco, J.; Neyman, K.; Illas, F. *Phys. Rev. B* **2007**, *75*, 035115.

(23) Da Silva, J. L. F.; Ganduglia-Pirovano, M. V.; Sauer, J.; Bayer, V.; Kresse, G. *Phys. Rev. B.* **2007**, *75*, 045121.

(24) Da Silva, J. L. F. *Phys. Rev. B* **2007**, *76*, 193108.

(25) Castleton, C. W. M.; Kullgren, J.; Hermansson, K. *J. Chem. Phys.* **2007**, *127*, 244704.

(26) Kresse, G.; Blaha, P.; Da Silva, J. L. F.; Ganduglia-Pirovano, M. V. *Phys. Rev. B* **2005**, *72*, 237101.

(27) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106.

(28) Hay, P. J.; Martin, R. L.; Uddin, J.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 034712.

(29) Ganduglia-Pirovano, M. V.; Da Silva, J. L. F.; Sauer, J. *Phys. Rev. Lett.* **2009**, *102*, 026101.

(30) Fabris, S.; de Gironcoli, S.; Baroni, S.; Vicario, G.; Balducci, G. *Phys. Rev. B* **2005**, *72*, 237102.

(31) Cococcioni, M.; de Gironcoli, S. *Phys. Rev. B* **2005**, *71*, 035105.

(32) Jiang, H.; Gomez-Abal, R. I.; Rinke, P.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 126403.

(33) Duclos, S. J.; Vohra, Y. K.; Ruoff, A. L.; Jayaraman, A.; Espinosa, G. P. *Phys. Rev. B* **1988**, *38*, 7755.

(34) Gerward, L.; Olsen, J. S. *Powder Diffr.* **1993**, *8*, 127.

(35) Branda, M. M.; Hernández, N. C.; Sanz, J. F.; Illas, F. *J. Phys. Chem. C* **2010**, *114*, 1934.

(36) Hernández, N. C.; Grau-Crespo, R.; de Leeuw, N. H.; Sanz, J. F. *Phys. Chem. Chem. Phys.* **2009**, *11*, 5246.

(37) Branda, M. M.; Castellani, N. J.; Grau-Crespo, R.; de Leeuw, N. H.; Cruz Hernandez, N.; Sanz, J. F.; Neyman, K. M.; Illas, F. *J. Chem. Phys.* **2009**, *131*, 94702.

(38) Kullgren, J.; Castleton, Ch. W. M.; Müller, C.; Muñoz-Ramo, D.; Hermansson, K. *J. Chem. Phys.* **2010**, *132*, 054110.

(39) Martin, R. L.; Illas, F. *Phys. Rev. Lett.* **1997**, *79*, 1539.

(40) Moreira, I. de P. R.; Illas, F.; Martin, R. L. *Phys. Rev. B* **2002**, *65*, 155102.

(41) Dovesi, R.; Saunders, V. R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Pascale, F.; Civarelli, B.; Doll, K.; Harrison, N. M.; Bush, I. J.; D'Arco, P.; Llunell, M. *CRYSTAL06 User's Manual*; Universita di Torino: Torino, 2006.

(42) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(43) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1998**, *298*, 113.

(44) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

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

(46) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1998**, *37*, 785.

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

(48) Bilc, D. I.; Orlando, R.; Rignanese, G. M.; Íñiguez, J.; Ghosez, P. *Phys. Rev. B* **2008**, *77*, 165107.

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

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

(51) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

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

(53) Wu, Z.; Cohen, R. E. *Phys. Rev. B* **2006**, *73*, 235116.

(54) Dolg, M.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1989**, *90*, 1730.

(55) The CRYSTAL web page: http://www.crystal.unito.it

(56) Corà, F. *Mol. Phys.* **2005**, *103*, 2483.

(57) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188.

(58) Pascale, F.; Zicovich-Wilson, C. M.; López Gejo, F.; Civarelli, B.; Orlando, R.; Dovesi, R. *J. Comput. Chem.* **2004**, *25*, 888.

(59) Saunders, V. R.; Freyria-Fava, C.; Dovesi, R.; Salasco, L.; Roetti, C. *Mol. Phys.* **1992**, *77*, 629.

(60) Sholl, D. S.; Steckel, J. A. In *Density Functional Theory: A Practical Introduction*; John Wiley and Sons: Hoboken, NJ, 2009, p 222.

(61) Marabelli, F.; Wachter, P. *Phys. Rev. B* **1987**, *36*, 1238.

(62) Wuilloud, E.; Delley, B.; Schneider, W. D.; Baer, Y. *Phys. Rev. Lett.* **1984**, *53*, 202.

(63) Lide, D. R., Ed. *CRC Handbook of Chemistry and Physics*, 9th ed.; CRC Press, Boca Raton, Florida, U.S.A., 2009).

(64) Muscat, J.; Wander, A.; Harrison, N. M. *Chem. Phys. Lett.* **2001**, *342*, 397.

CT100430Q

JCTC Journal of Chemical Theory and Computation

# Designing a Scalable Fault Tolerance Model for High Performance Computational Chemistry: A Case Study with Coupled Cluster Perturbative Triples

Hubertus J. J. van Dam,* Abhinav Vishnu,* and Wibe A. de Jong*

*Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99354-1793, United States*

Received August 6, 2010

**Abstract:** In the past couple of decades, the massive computational power provided by the most modern supercomputers has resulted in simulation of higher-order computational chemistry methods, previously considered intractable. As the system sizes continue to increase, the computational chemistry domain continues to escalate this trend using parallel computing with programming models such as Message Passing Interface (MPI) and Partitioned Global Address Space (PGAS) programming models such as Global Arrays. The ever increasing scale of these supercomputers comes at a cost of reduced Mean Time Between Failures (MTBF), currently on the order of days and projected to be on the order of hours for upcoming extreme scale systems. While traditional disk-based check pointing methods are ubiquitous for storing intermediate solutions, they suffer from high overhead of writing and recovering from checkpoints. In practice, checkpointing itself often brings the system down. Clearly, methods beyond checkpointing are imperative to handling the aggravating issue of reducing MTBF. In this paper, we address this challenge by designing and implementing an efficient fault tolerant version of the Coupled Cluster (CC) method with NWChem, using in-memory data redundancy. We present the challenges associated with our design, including an efficient data storage model, maintenance of at least one consistent data copy, and the recovery process. Our performance evaluation without faults shows that the current design exhibits a small overhead. In the presence of a simulated fault, the proposed design incurs negligible overhead in comparison to the state of the art implementation without faults.

## 1. Introduction

In the past couple of decades, computational chemistry has developed to the point where it has become a significant tool in solving real world problems. The development has been supported by a tremendous growth in computational power provided by the most modern supercomputers. While this growth initially was delivered by increasing single processor capability, the past two decades have observed a significant part of this growth through the development of parallel computing systems. This development has recently been accelerated by the arrival of multicore and GPU-based systems. Currently, the extreme end of this development is embodied in the Jaguar machine,[1] which employs 224 256 cores, delivering at most $2.3 \times 10^{15}$ floating point operations per second.[2] Presently, researchers are looking into the possibility of 1000 times more powerful supercomputers.[3]

At the same time, there has been a development toward parallel implementations of computational chemistry applications to take advantage of these supercomputers. A number of parallel quantum chemistry applications have become available, including NWChem,[4] GAMESS,[5] GAMESS-UK,[6] MOLPRO,[7] MOLCAS,[8] Q-Chem,[9] PQS,[10] MPQC,[11] ADF,[12] Dalton,[13] FreeON,[14] and COLUMBUS,[15] to name a few. The different applications and different methods

* Corresponding authors. Phone: +1 509 372 6441. Fax: +123 (0)123 4445557. E-mail: Hubertus.vanDam@pnl.gov (H.J.J.v.D.); Abhinav.Vishnu@pnl.gov (A.V.); Wibe.deJong@pnl.gov (W.A.d.J.).

10.1021/ct100439u 
Published on Web 12/08/2010

implemented by these applications show different characteristics in the parallelism they can exploit. Ongoing developments continue to improve these capabilities, and chemistry continues to be at the forefront of parallel applications, as recently demonstrated by Apra et al.[16]

However, the increasing scale of these supercomputers has reduced the Mean Time Between Failures (MTBF) sharply. The current generation of supercomputers exhibits an MTBF of 6 h,[17] seriously limiting the overall system usage of these supercomputers. To deal with such problems, disk-based checkpoint/restart fault recovery strategies have usually been implemented. In this approach, the state of an application is written to disk at regular intervals. After a fault occurs, the last valid state is restored and the calculation continues from that point. This mechanism can be provided by the application itself writing restart files at particular points during a calculation or by the operating system saving the state of the calculation in an application-transparent manner. The problem with the checkpoint/restart approach is that the I/O overhead it introduces is significant. In fact, it is estimated that, somewhere between peta- and exa-scale calculations, the time spent on the overhead of the checkpoint/restart fault tolerance strategy will exceed the time spent doing useful computation.[18] In addition, quite a few times, the act of checkpointing itself brings the overall system to a halt. Clearly, other fault tolerance strategies for efficient recovery are imperative for computational chemistry methods. This leads to the following challenges:

• designing fault tolerant methods that allow computational chemistry codes to proceed correctly in the presence of faults
• ensuring that the overheads introduced are insignificant

These challenges were considered by Nielsen et al.[19] to some extent, but they judged implementing a fault tolerant method impractical for the lack of fault tolerance support in MPI. In this paper, we address these challenges by designing a fault tolerant strategy based on in-memory data redundancy using Global Arrays (GA).[20] GA uses a message passing layer, but it can use its own TCGMSG[21] layer rather than MPI. Although we have no control over MPI, we can design a fault tolerant GA/TCGMSG infrastructure as the foundation for fault tolerant applications. The aim of this approach is to develop applications that can withstand system failures while incurring low overheads. While the approach is not specific to chemistry applications, the impact is related to algorithm-specific parameters such as data volumes, message frequencies and sizes, communication/work balance, etc. Therefore, it is relevant to assess the appropriateness of the chosen approach to the target applications. To this end, we have implemented a fault tolerant version of the coupled cluster perturbative triples. We assess the level of fault tolerance that can be attained, the extent of the code modifications needed, as well as the overheads incurred.

The rest of the article is organized as follows. In section 2, we present the background of our work. In section 3, we present the design of our work. Section 4 presents the performance evaluation of the fault tolerant coupled cluster method. In section 5, we present the conclusions and future directions. We begin with the description of the background.

## 2. Background

The increasing scale of modern supercomputers has not only reduced the MTBF, it has also increased the variety of faults which occur during the execution of the applications. While disk failures continue to decrease due to the arrival of diskless systems (Cray XT5 and IBM Bluegene), processor failures due to overheating, node failures due to power tripping, and network failures for high connectivity interconnects are ubiquitous. Some of these failures lead to a permanent change of the calculation, e.g., a node failure due to power tripping leads to the permanent loss of the processes running on that node. These kinds of failures are referred to as hard failures.

Other failures, such as certain kinds of random memory bit flips as a result of radiation,[17] are transient or intermittent. They do not occur at a specific time and may not be reproducible at a later stage. These kinds of failures are referred to as soft failures. They are also inherently harder to detect because if such a fault is suspected there is no general *a posteriori* test that can reliably establish that such a failure actually has happened. The only way to reliably test for such faults is through redundancy, either in hardware, e.g., extra bits used in Error Correction Code (ECC),[22] or software by redundant execution of instructions and checking that the results are the same.[23]

During most extreme scale application executions, combinations of these failures are expected to occur. In the present work, we target only hard failures. Moreover, we assume that most hard failures will be detected by the runtime system and result in the affected processes being terminated. In particular, we assume that in most cases the impact will be such that all processes on a particular node will be terminated. Obviously this does not mean that soft failures are insignificant. However, provided the capability to detect soft failures is available, it is always possible to render soft failures into hard failures by terminating the affected processes.

**2.1. Fault Behavior and General Data Fault Resilience.** Given that a fault expresses itself in the termination of one or more processes, there are generally three different possible choices. The most commonly supported choice at the moment is to restart the application on the basis of a checkpoint/restart mechanism. As mentioned in the Introduction, the overhead of this approach is high and expected to become impractical for the scale of calculation currently envisioned.

Another approach is to detect the process failures and recover by reinstantiating the failed processes. In the case of processes failing due to hardware becoming unavailable, e.g., a power supply failing, effectively turning a number of processors off, the processes would have to be reinstantiated on additional resources. These resources would have to be on stand-by for as long as no fault occurs, which is wasteful. However, more problematic is that upon reinstantiating the processes they would have to be put in a state that is consistent with the other processes in the calculation. For many chemistry applications, this is highly nontrivial, as in normal circumstances, the state is updated in the course of the calculation. Hence, the state is a function of the program and the path its execution has taken through it up to a

particular point. This path often involves the modification of many variables that manage the calculation to be executed, such as I/O, communication, data distributions, etc. The application relies on most of these variables at a later point in the calculation. Hardly ever is the relationship between the values of these variables and the particular phase of the computation in hand explicitly formulated. Therefore, recreating a process and bringing it to the correct state is at best hard to achieve.

The remaining approach is to accept the loss of the failed processes and continue the calculation without them. Obviously, this leads to a degradation of performance, but as long as this is roughly proportional to the fraction of processes lost, this is acceptable. In most cases, the loss of performance has less of an impact than the interruption caused by checkpoint/restarting or forcing an application shutdown. In extreme cases where the loss of performance is severe, this essentially points to a serious instability of the machine, and continued execution is not recommended anyway.

Considering the implications of the various approaches, we consider the last option of continued execution in the degraded mode the most promising option. In designing such an approach, an important question is how to handle the impact a fault has on the data in the application. Clearly, as a fault leads to the loss of processes, the data those processes hold become inaccessible to the application. Broadly speaking, there are two possible ways to address this issue (disregarding keeping copies on disk, which would effectively be equivalent to checkpoint/restart).

One way to address this is to replace the lost data with "reasonable" guesses for the lost data. If the guesses are close enough and the amount of data lost small enough, one could expect the calculation to continue almost as normal with only a slight perturbation. This approach could work for optimization algorithms where the end-point is specified in terms of certain conditions being met. In this case, a small upset along the way might not have significant consequences. The main flaw in this idea is however that many data structures have to satisfy specific conditions. For example, molecular orbitals have to be orthonormal, the trace of the density matrix has to equal the number of electrons, wave functions have to be normalized, etc. Just replacing a patch of the data with some guess is more than likely going to violate these conditions. Explicitly re-establishing these conditions is often expensive and involves a global response; think for example of reorthonormalizing the molecular orbitals. The alternative would be to write the algorithm in such a way that it intrinsically strives to satisfy these conditions; i.e., meeting these conditions to a certain degree becomes part of the convergence criterion. In the abstract, it is not clear what the computational cost implications of such an approach are, but it is clear that this would require nontrivial modifications of all algorithms to which it is applied. Even then, this approach could only be used for optimization algorithms. It offers no way forward, for example, for algorithms like the MP2 energy evaluation, which is a single step method. In those cases, any perturbed input data will lead to a perturbed answer.

The other approach is to rely on redundant data of some kind. Obviously, disk-based data redundancy is out of question due to the I/O overheads, but in-memory redundancy is an option. In particular, in quantum chemistry where the computational complexity is always of a higher order than the data storage requirements (only in the large systems limit for linear scaling methods both orders become the same), it is possible to duplicate the critical data. The assumption is that the choice of the number of processors is motivated by the amount of work to be done rather than the amount of memory needed. This is certainly true for correlated quantum chemistry methods. The main cost of this approach will be in the extra communication required to maintain the duplicate copies. Provided the granularity of the work is high enough, this should not be a major problem.

In this paper, we describe how fault tolerance can be achieved using in-memory data redundancy in the context of degraded mode continued execution.

**2.2. Global Address Space Programming Models and Global Arrays.** As will become clear in the next section, a critical capability for our fault tolerance approach is that other processes can continue from the point where some process failed. Central to this capability is access to the data required to do so. Although in principle this access can be arranged in a number of ways and environments, in practice it is particularly facilitated by the Partitioned Global Address Space paradigm (PGAS). In this paradigm, physically distributed objects can be created, but the access mechanisms make it appear as if they exist in a shared memory space. This also means that every process can access all of each such object independently of what other processes are doing. Today, this programming model is supported by a number of programming languages such as Unified Parallel C[24] and Co-Array Fortran.[25]

However, before these programming languages were defined, it was already realized that the PGAS paradigm can be very helpful in solving some problems. In particular, in the context of the development of NWChem, this paradigm offered a way to deal with collections of tasks with a large spread in execution time without incurring large load imbalance overheads. To support this paradigm, the Global Array toolkit (GA) was created.[20] This library provides an efficient and portable "shared-memory" programming interface for distributed-memory computers. Each process in a Multiple Instruction Multiple Data (MIMD) parallel program can asynchronously access logical blocks of physically distributed dense multidimensional arrays, without the need for explicit co-operation by other processes. Unlike other shared-memory environments, the GA model exposes to the programmer the nonuniform memory access (NUMA) characteristics of high-performance computers and acknowledges that access to a remote portion of the shared data is slower than to the local portion. The locality information for the shared data is available and can be exploited to maximize the performance of codes if desired.[20]

Combining the global data access characteristics provided by the GA with data redundancy, it becomes particularly easy to ensure access to all data even in the case where processes fail. Hence, this library provides a highly suitable

platform with which to construct fault tolerant applications. Obviously, when the GAs were first written, fault tolerance was not envisioned as a problem. Therefore, changes to the library itself are required to make it fault-resilient. These changes are nontrivial, as they touch every layer in the library's software stack. They include considering how the library should respond to basic communication calls involving failed processes, implementing collectives such as barriers with incomplete sets of processes, as well as fault detection mechanisms to establish which processes have been lost and propagating that information to the processes that need to know, and modifying the process manager (the equivalent of mpirun for MPI applications) to respond correctly to the loss of processes. This work is currently ongoing, but its discussion is beyond the scope of this paper. Instead, in this paper, we limit ourselves to describing how a fault-resilient GA implementation can be used to build fault tolerant computational chemistry applications. For performance evaluations, a traditional GA library is used, and faults are simulated by making selected processes behave as if they have failed. This triggers the same responses in the application and the fault tolerance layer as a real fault. Therefore, it is expected to give a realistic impression of the behavior of the real fault-tolerant code.

**2.3. The Coupled Cluster Method.** The Coupled Cluster (CC) method has since its inception[26] developed into the dominant method for highly accurate calculations of total energies. Due to the relatively high cost of the method, a hierarchy of methods has been developed of increasing accuracy and cost. The natural relationship between CC and perturbation theory allows for the addition of perturbative corrections that offer increased accuracy at a lower cost than the full CC method at the corresponding level of approximation. The most widely used method of this kind is the CCSD(T) method. This refers to a CC method in which the singly and doubly excited Slater determinants are accounted for in the iterative solution of the equations, the CCSD part. The triply excited Slater determinants are accounted for using perturbation theory, which is indicated with (T).

The theory of CC methods has been described in a number of papers.[26−29] In this paper, we concentrate on creating a fault-tolerant implementation of the approach by Kobayashi and Rendell.[30] The initial implementation is characterized by an iterative CCSD solver that has a cost that scales as $O(N^6)$. Every iteration, this solver writes out the current values of the amplitudes, which allows this part to be restarted. The CCSD part is followed by the (T) term to form CCSD(T). This part requires transformed integrals with at most three virtual indices, which cost $O(N^5)$ to compute, followed by the triples contributions that cost $O(N^7)$. Because the triples contributions are calculated perturbatively, they are evaluated in a single pass. Although at $O(N^7)$ the triples contributions are the most expensive part because they are evaluated in a single pass, there is no natural point from which to restart their evaluation.

Despite the algorithm having no natural restarting point, this does not mean that it is impossible to make the code restartable. In doing so, the transformed two-electron integrals would have to be recalculated, as storing them would require excessive amounts of disk space. For the triples contributions themselves, the total number of tasks could be divided into blocks of some fixed number of tasks. At the end of each block, the current total triples contribution could be saved along with the task number of the next task. This would provide the data needed for restarts. The disadvantages of doing this are increasing the load imbalance due to extra barriers needed at the end of every block of tasks and a loss of productivity, as restarting a calculation requires resubmitting the job and waiting for its execution. Hence, having a facility that allows the code to continue without restarting is usually beneficial.

In addition to the fact that it is beneficial not to have to restart the perturbative triples, their high cost and good scalability make this code one of the prime candidates to be executed on extreme scale computing platforms. Indeed, this code was run on 200 000+ cores on Jaguar already.[16] So this code is likely to run in cases where fault tolerance is essential and therefore a suitable target application for our approach. In the future, we plan to extend our fault tolerance approach to other code modules as well.

## 3. Overall Design

In the following sections, the model for achieving fault tolerance based on in-memory redundancy is outlined. Important questions pertaining to this model are

- What kind of failure scenarios can this model address?
- How extensive are the code changes needed to integrate this model into an application?
- What are the performance overheads if no faults occur?
- What is the overhead to recover from a fault? To answer these questions, we built a fault-tolerant version of CCSD(T) using the model described here. With this implementation, we illustrate the answers to these questions.

A general aspect in a redundancy-based fault tolerance model is that when a process fails its memory becomes inaccessible, and the data it held are lost. Therefore, these data need to subsequently be fetched from a different location. To facilitate these changes in the communication pattern, it is essential to have a memory management approach that is sufficiently flexible. The GA[31] provides a virtual shared memory programming model that gives every process the same view of a distributed data structure. This way, it offers sufficient flexibility for the approach we propose. Therefore, on the application side, we have built an infrastructure on top of the GA that manages the application response to faults, as outlined below. We have currently integrated this infrastructure with NWChem[32] to deploy it in real chemistry applications.

NWChem breaks large-scale calculations up into small tasks that work on particular blocks of data. When a failure occurs, the responsibility for a particular task needs to be transferred to another process. To enable this, visibility of a particular process's work to other processes has to be ensured.

Combining the knowledge of a process's state and the redundancy of the data, it is possible to devise a recovery mechanism that returns the overall calculation to a valid state after a fault has occurred. In the following subsections, we

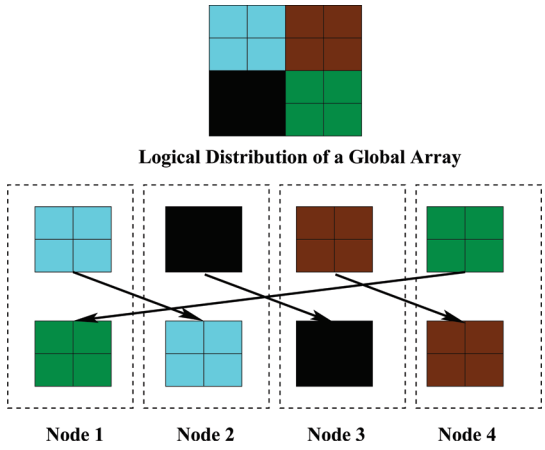



***Figure 1.*** Relationships between the logical array, the physically distributed primary copy, and the distributed shadow copy.

present the details of the data storage model, the task model, and the fault recovery mechanism. This description of the general approach is followed by a description of how this is applied in the triples part of the CCSD(T) code.

**3.1. The Data Storage Model.** As stated in the outline of the general approach, data redundancy is used to enable fault recovery. This means that for every distributed data object a primary copy is created that is identical to a GA as usual. Figure 1 shows, at the top, the logical view of the GA. This array is partitioned into blocks across a processor grid. The choice of the processor grid depends on the number of dimensions in the GA, the sizes of the dimensions, and the number of processors. In practice, a single process is associated to a given "processor" in the processor grid. After this partitioning, every process is given at most a single data block of the GA, as shown in the middle.

In addition, a shadow copy is created. The data in the shadow copy are shifted by a given number of processes relative to the primary copy in a cyclic fashion, as shown at the bottom. The shift is referred to as the "processor shift" and can be chosen arbitrarily as long as it is not an integer multiple of the number of processes (that would lead to both the primary and shadow copies of the data residing at the same process). Obviously, under the assumption that all processes on a single node are likely to fail together, the processor shift should be at least equal to the number of processes per node to ensure that the shadow copy resides on a different node than the primary copy.

The processor shift is implemented using the new restricted array capabilities of the global arrays. This allows for physical data blocks to be mapped to processes using an arbitrary mapping. Hence, all of the administration related to the mapping between the primary and shadow copies can be off-loaded to the GA. In addition, the restricted array mappings also allow mappings that are not simply a cyclic shift. For a given machine, this capability can be exploited to place the data such that the application fault tolerance is maximized while minimizing the communication cost. Doing this requires knowledge of the anticipated fault behavior of the machine as well as the topology of the interconnect. At present, we do not exploit this, but in the future we could do so if that turns out be beneficial.

If being able to access data after a failure was the only concern, it would be sufficient to have only the primary and shadow copies. In practice, however, it is possible for a process to fail while it is updating data held on remote processes. This data will still be accessible after the failure but its contents might be corrupted. Therefore, a facility is needed to record the status of every data block. Hence, for both the primary and shadow copy, an additional GA is needed with a length of the number of processes. These arrays serve as flags to signal whether the corresponding data block at the particular process is "clean" or "dirty".

In summary, in this approach, every GA in the original implementation is replaced with a redundant GA for fault tolerance. Every redundant GA consists of four GAs, the primary copy, the shadow copy, the primary copy status flags, and the shadow copy status flags. In our implementation, this collection of data structures is treated as a single object with an interface that is essentially the same as the one for the usual GAs.

In analogy to the usual GA implementation, the interface to manage these data structures implements five routines:

(1) nsft_create: This creates an N-dimensional distributed array with its associated redundant copies and the block status flags.
(2) nsft_destroy: This clears a distributed array and its associated data structures up.
(3) nsft_map_put: This stores data in a distributed array, managing access to the shadow copies and the status flags. It can transfer data to multiple arrays in a single call.
(4) nsft_map_acc: This is the same as nsft_map_put apart from the fact that, whereas put overwrites the original data, the accumulate (named "acc" for short) routine adds to it.
(5) nsft_get: This gets a patch of data from any of the copies that are available.

The put and accumulate routines are for the most part straight equivalents of the corresponding routines in the GA library. However, the need to be able to update multiple arrays in one call is a requirement of the fault recovery, and the details of this will be discussed below.

**3.2. The Task Model.** Traditionally, the work in a given compute phase is broken up into tasks. Any task in some way involves

- establishing which task to do
- retrieving the data needed to complete the task
- executing some instructions to generate results
- storing the results in the appropriate places

In practice, these steps do not necessarily have to be executed in this order. For example, a task may get some data, do a bit of work, store some results, get some more data, do more work, and accumulate the result in a local buffer (the buffer could be globally summed after all tasks have been completed). However, if the calculation has to be able to recover from faults, a stricter set of rules is needed.

First of all, in order to have a simple way to identify what every process is doing, every task is associated with a unique number within a particular computational phase. This task number dictates what work is involved in the task, what data

it needs, and, most importantly, what data it is going to change. The advantage of this is that if a task needs to be re-executed due to a fault, the recovery mechanism only has to pass the task number to a suitable process to identify what needs to be done. To some extent, task numbers are already implicitly used in this way but often with the assumption that the task numbers are monotonically increasing. In a fault-tolerant application, this assumption no longer holds, as after a fault, the recovery mechanism may pass an old task number to a process for re-execution.

So as a general rule, the task model will insist that a process loops over tasks. The tasks may come in a random order, and whenever a task number is in the valid range from 1 to the number of tasks, the task parameters are calculated from the task number, and the required work is done. If the task number is outside the valid range, this means that all tasks have been done. Typically, the process then proceeds to a barrier that marks the end of the compute phase in question.

As far as accessing data is concerned, the fault-tolerant task model imposes extra rules only on the way data are changed. As a general rule, fault tolerance requires that when a fault occurs the state of a calculation can be either restored to the point where it is as if a task was never started or progressed to the point where it is as if a task was fully completed. This all or nothing approach is very much akin to transactional systems.[33] Transactional approaches are commonly used in distributed databases and business administration systems. They have two characteristics that are crucial to these application areas: atomicity and serializability. Atomicity in this context means that all modifications of all data items in a transaction are either progressed to completion or no updates happen at all, leaving all persistent data structures in their original state. Serializability means that all transactions complete as if they were executed in serial in some order. This implies that no process can get a view of the data where only part of the transaction has taken place, i.e., any of the intermediate states of the transaction. The implementation of these approaches rely on a two-stage mechanism for executing transactions. In the first phase, the data needed for the updates are sent out to all of the relevant data servers, but the persistent state is not changed. In a second phase, the changes are committed, meaning that the changed data structures take effect in the persistent state and the previous versions are discarded. If for some reason not all servers can commit, an abort can be issued, causing all modified versions to be discarded and the unmodified versions to remain part of the persistent state. Crucial is the separation between the data transmission and changing the persistent state of the data.

Although, in principle, elements of transactional updates could be implemented in the GA, at present such mechanisms are not available. The reasons for not supporting these mechanisms are related to both need and cost. If fault tolerance is not an issue, then there is no benefit in having a transactional mechanism. To the contrary, because on the data server side the information flows through an additional copy, three memory accesses are needed for each data element. In the current implementation, only one memory access is needed when the data are directly received into the persistent data set. In addition, the transactional approach requires nontrivial amounts of extra memory as well for storage of data between transmission and committing.

In our approach, the basic GA functionality is unchanged. This, however, requires that all primary copies are updated at the same time, so that if a fault occurs, all of those blocks can be marked dirty, and the unchanged shadow copy blocks count as the true record of the data. It also requires that all of the shadow copies be updated at the same time, so that if a fault occurs at that stage, all affected shadow copy blocks can be marked dirty, and the already updated primary copies count as the true record of the data. Obviously, a scenario in which some arrays have both their primary and shadow copies updated whereas others were left unchanged altogether leaves the calculation in an unrecoverable state. To facilitate writing code that adheres to these data update requirements, the map-based put and accumulate routines are provided. These routines can update as many arrays as needed in one call and ensure that the updates are executed in a way that allows for fault recovery. To ensure correctness, however, we also have to enforce that there can only be at most one call to update data per task. Finally, in order for a different process to continue from where a failed process left off, it is essential that every task stores its results in a distributed data array. No results shall be stored in local buffers beyond the scope of a task.

Obviously, for the fault recovery mechanism to decide on the right course of action, it needs to know what a process was doing at the time of failure. Of course, it needs to know which task was being executed, but it also needs to know what was being done on the task. In particular, we distinguish four different states:

(1) acquiring the task number
(2) working (getting data and computing results)
(3) updating primary data copies
(4) updating shadow data copies

These task states together with the task number are stored in a distributed and redundant GA with one element per process, so that the recovery mechanism can access this information when needed. The management of these states can be completely hidden from the programmer.

The task model needed for fault tolerance is summarized in Figure 2. The two gray sections are provided by the infrastructure for fault tolerance, whereas the white section is application-dependent code. The arrows indicate additional communication to record the state of a task in a globally accessible place. This way, the task state can be recovered in case the task fails.

**3.3. The Fault Recovery Mechanism.** The components discussed so far are essential ingredients in making fault tolerance possible. The fault recovery mechanism is the part that makes it work. It has to address the following questions:

- who executes the recovery mechanism?
- when is the recovery mechanism invoked?
- what steps are taken?

To answer the question of who is to respond to a fault, every process is assigned a "buddy" process. In this context, the two processes in a buddy pair are inequivalent in that

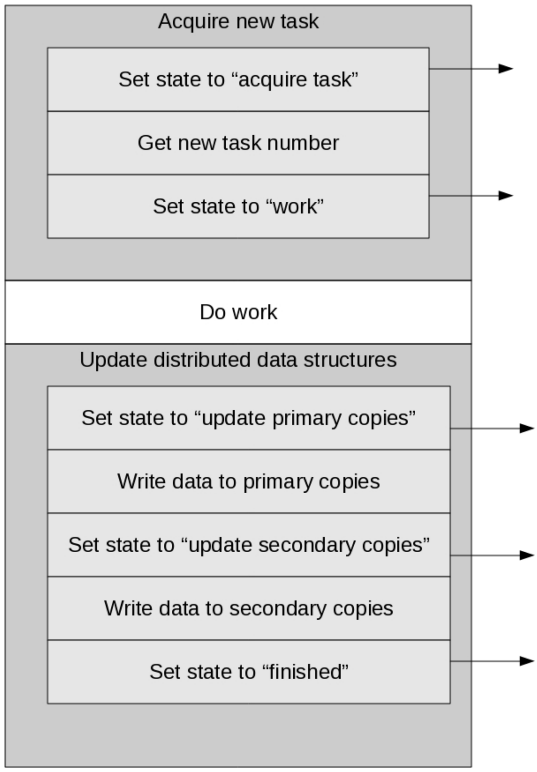



***Figure 2.*** Outline of the stages in the fault-tolerant task model. The arrows indicate additional communication to record the task state changes needed.

the buddy of a process steps in and takes care of the process's responsibilities when it no longer can due to a fault. However, this relationship does not work the other way around. For a given process, its buddy is chosen by subtracting the processor shift (as used in the data storage scheme) of the process rank.

The most convenient time for the buddy process to respond is when it reaches its next time to acquire a new task number. This allows the routine that generates the new task number to take generic remedial steps, as outlined below, and if the failed task needs to be re-executed, its number can be passed to the buddy process as its next task. This way, the remedial steps can be almost completely hidden from the programmer, and no code has to be written to specifically deal with the re-execution of tasks. Instead, the normal code can be used for recovery-driven re-execution as well as normal tasks.

The recovery process follows a very simple approach to ensure that the data the program uses are in a consistent state and that, where necessary, tasks that failed are finished. The precise steps that need to be taken depend on the state of the tasks of the failed process at the time the fault occurred. Obviously, the buddy process first has to retrieve the task number and state of the failed process, which are stored in a redundant GA. If the state of the task was "acquiring new task", at present, recovery is not possible, and the calculation terminates. As shown in section 3.4, this situation is very unlikely to arise, but if it becomes a problem, it can be addressed by recording the status of every task rather than every process. If the state was "working", then no remedial action is needed, but the failed task needs to be re-executed. If the state was "updating primary copies", then the contents of the associated memory are undefined. Hence, the memory blocks associated with the failed task are worked out, and those blocks are marked "dirty". The programmer has to provide a routine to the task generator that can provide a list of blocks a task will change on the basis of the task number and process rank. As this relationship is highly dependent on the particular algorithm the program executes, there is no general other way of obtaining that information. If the primary copies were corrupted, the task results will not have been stored in the shadow copies either, and the task needs to be re-executed. Finally, if the state was "updating shadow copies" then the associated blocks of the shadow copies are marked "dirty". However, as the results are already stored in the primary copies, there is no need to re-execute the task in this case.

Concluding this section, it is clear that most of the details of the fault recovery can be hidden from the programmer. The only thing that cannot be hidden is the relationship between the task number and the data that task updates, as this is an intrinsic property of the application algorithm. However, generating this dependency should be trivial for most algorithms.

**3.4. Summary of the Fault Tolerance Approach.** Considering the approach outlined above, it is clear that this approach cannot deal with all faults. The approach explicitly assumes that faults lead to process termination and that soft faults do not occur. In addition, the failure of two processes differing in rank by the processor shift destroys both the primary and shadow copies of some of the data, leaving the calculation in an unrecoverable state. However, the likelihood of such failures is the square of the probability of any given process failing. That is assuming that the processor shift is chosen such that the likelihood of the one node failing is independent of a failure of the other. Hence, this should be very small for any reasonable machine.

Furthermore, at present, the approach cannot recover from a fault occurring when a process is acquiring the next task number. The reason is that if the global counter generated a task number but the receiving task failed before storing this number in a globally accessible place, then this task is lost. At present, this is not expected to be a significant problem, as the code spends very little time establishing the next task number compared to the rest that it does. So the chance of failing in this phase is very small. However, if needed there are at least two ways in which the problem can be addressed. One is to change the global counter to keep a record of the last task number it handed out to each process. Storing this record in a GA keeping all data elements on the same process as the global counter itself allows an efficient implementation. The reason is that the data are written often but read once. That is, the data only need to be read if a task failed while getting the next task number. Then, checking the consistency between the task state and the last global counter generated task number allows the determination of what task should be executed. The other way to address this issue is to store the state of every task. At the end of the compute phase, all tasks should be completed. Ones that are not must have remained uncompleted due to faults, and remedial actions can be triggered to complete them nevertheless.

Finally, the recovery process can cause a loss of data redundancy. If a process fails while updating the primary or shadow copies, the corresponding copy is marked "dirty", after which only the other copy remains. It is tempting at that stage to try and salvage the dirty copy with data from the clean copy. However, during the compute phase, this can lead to inconsistencies. To see this, consider processes A and B, which are executing tasks that are both going to update the same patch of a redundant GA. The primary copy of that patch resides on process P and the shadow copy on process S. Now process A fails while updating P, so the data are P is dirty. Process B has finished its task, updates P, and is about to update S. At this point, the salvage procedure kicks in copying the data from S to P and marking P "clean" again. Subsequently, process B updates S. At this time, the consistency between the redundant copies is lost. The primary copy contains data without the contribution from B, and the secondary copy includes the contribution from B. A similar scenario can cause problems attempting to salvage a "dirty" shadow copy by using the data from the primary copy. So these salvage operations should not be attempted during the compute phase when a number of processes could simultaneously be changing the redundant GAs. The only time such salvaging can be done is as part of the barrier at the end of a computational phase. Only at that time can one guarantee that copying data between the primary and shadow copies does not lead to inconsistencies due to missed updates.

The loss of redundancy due to dirty blocks is obviously a direct consequence from the way the GA implements data updates. If a transactional GA interface were implemented and used, this issue would not arise. Depending on how severe this issue becomes, this might be a reason to develop a transactional approach.

**3.5. Code Changes Required for Fault Tolerance.** Aside from the question of which faults this approach can deal with, there is the question about the extent of the code changes needed to implement it. In Figure 3, a section of pseudocode is shown that illustrates the changes needed to implement fault tolerance in the CCSD(T) code. The first obvious difference is how in the fault-tolerant version key task parameters are derived from the task counter. In the original version, they were established using an implicit dependency on the task counter, but as discussed that dependency is broken due to the fault tolerance. Second, fault tolerance requires that intermediate results are kept in a globally accessible space. Hence, the local variable "sum" had to be replaced by a redundant GA. The contributions to the energy are added on by the sft_map_acc routine, which applies consistent updates to the primary and shadow copies. In this case, it only adds to one array, but it can modify any number of arrays. Third, notice the additional parameter to sft_nxtask, which is a routine that informs the recovery process about the data blocks a given task number modifies. As this information is application-specific, this routine needs to be provided by the application programmer. Finally, after all contributions have been summed into "g_sum", the various partial results need to be collected and added together to arrive at the final answer. Obviously, some code changes are needed, and some thought needs to go into this to arrive at a fault-tolerant implementation. Nevertheless, the example clearly shows that these changes are relatively minor. In particular, the administration involved in maintaining a calculation in a recoverable state can be completely hidden from the application programmer.

```
itask = -1
sum = 0.0
next = nxtask()
do a = 1, nvirt
  do j = 1, nocc
    itask = itask + 1
    if (itask.eq.next) then
      call ga_get(g_in,1,nocc,
                  a,a,d,len)
      ... do work ...
      sum = sum + term
      next = nxtask()
    endif
  enddo
enddo
call ga_dgop(sum)
```

```
call sft_create(g_sum,1,nproc)
call sft_zero(g_sum)
numtasks = nvirt*nocc
next = sft_nxtask(access_map)
do while (next.le.numtasks)
  a = next/nocc+1
  j = next - (a-1)*nocc
  call sft_get(g_in,1,nocc,
               a,a,d,len)
  ... do work ...
  call sft_map_init(map)
  call sft_map_add(map,g_sum,
                   term)
  call sft_map_acc(map)
  next = sft_nxtask(access_map)
enddo
call ga_sync
sum = 0.0
do ip = 1, nproc
    call sft_get(g_sum,1,1,
                 ip,ip,term)
    sum = sum + term
enddo
call sft_destroy(g_sum)
```

***Figure 3.*** Comparison of normal code versus fault-tolerant code.

## 4. Performance Evaluation

The maintenance of multiple copies of distributed data structures and the task states all come at a cost. Therefore, an important question is whether the approach suggested is actually practical or whether the overheads make it unusable. To assess these overheads, we performed calculations on uracil with a cc-PVTZ basis set[34] using Cartesian basis functions. The resulting calculation involves 340 orbitals, of which 29 are doubly occupied. All orbitals were correlated in the CCSD calculation. Subsequently, we performed CCSD(T) calculations both with the original code and with the fault-tolerant code.

The calculations were run on Chinook.[35] Chinook is a 160 TFLOP that consists of 2310 HP DL185 nodes with dual socket, 64-bit, Quad-core AMD 2.2 GHz Barcelona Processors. Each node has 32 GB of memory and 365 GB of local disk space. Communication between the nodes is performed using InfiniBand with Voltaire[36] switches and Mellanox[37] adapters. The system runs a version of Linux based on Red Hat Linux Advanced Server. A global 297 TB SFS file system is available to all of the nodes.

The calculations were run on three different processor counts, i.e., 256, 512, and 1024 processors. As the calculation is relatively small, it did not scale well to 2048 processors. For each processor count, the calculations were run three times, each time as a new job so that different node pools were used. Although this does not provide a statistically significant sample, it allows spurious results to be ruled out. The timings obtained were averaged over the three runs and

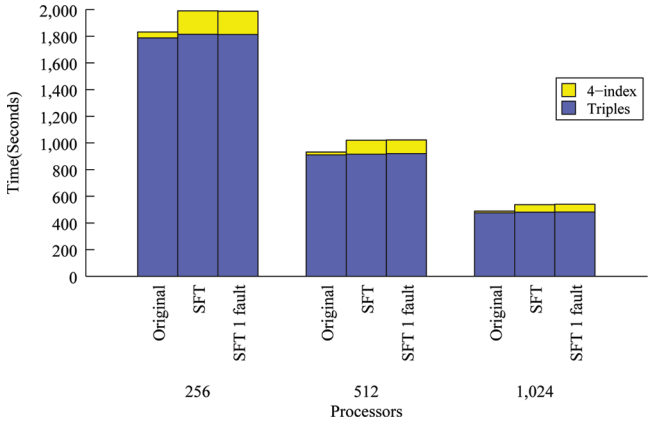



***Figure 4.*** Performance evaluation, on 256, 512, and 1024 processors comparing the original code, the fault-tolerant code (SFT), and the fault-tolerant code with a single fault (SFT 1 fault). The total time is split into the four-index integral transformation and the triples energy contributions.

are displayed in Figure 4. The figure compares for each calculation the time taken by the original code, the fault-tolerant code without faults, and the fault-tolerant code with one fault at the beginning of the fault-tolerant code section. The results show that the overhead incurred is 8.7%, 9.5%, and 9.9% for 256, 512, and 1024 processors, respectively. This overhead is largely incurred during the integral transformation phase of the calculation; the triples evaluation incurs an overhead of only 1.5%. As the triples component has a complexity of $O(N^7)$ and the integral transformation $O(N^5)$, this problem becomes less severe for larger calculations. In addition, the overheads are relatively insensitive to the processor count. Hence, when no faults occur, the overheads are sufficiently low for this technology to be used in routine calculations. We also measured the fraction of the time spent obtaining new task numbers, for which we found that 0.13% of the time the code is in a state where it currently cannot recover from a fault. Hence, in the vast majority of fault instances, the code will be able to continue with the current implementation.

So far, the case where no faults occur was considered, but the overhead of recovering from a fault is very important for practical applications. However, establishing this overhead is slightly complicated by the fact that, by continuing in degraded performance mode, any measurements will mix two aspects. Initially, there is the cost of bringing the data back into a valid state plus potentially re-executing the failed task. After that, there is the impact of continuing the calculation with one less processor. In practice, the timings fluctuate too much to separate these two effects. Hence, we simulate a process failing on its first task, as this will have the maximum impact and thus provide an upper limit.

The results in Figure 4 show that the loss of a single process leads to a performance degradation of at most 0.71% where one processor corresponds to 0.39% of the total computing capacity with 256 processors. This performance degradation includes the initial execution of the failed task, the fault recovery, the re-execution of the failed task, and the loss of one processor's compute capacity for the remainder of the calculation. Overall, this performance degradation is on the same order of magnitude as running the calculation on one processor less throughout; this impact is definitely small compared to the impact of aborting the whole calculation.

## 5. Conclusions

In this paper, we have discussed the challenges posed to computational chemistry by the fact that extreme scale computers are likely to have relatively small MTBFs. To address this, we have considered making real computational chemistry code fault tolerant. The aim of this effort is to allow the program to correctly continue execution despite the loss of processes. We have presented an approach that offers this capability, integrated it into NWChem, and evaluated the performance of the method for the perturbative triples component of CCSD(T) calculations. It was shown that the fault-tolerant code runs with an overhead of only a few percent relative to the original code. The occurrence of a fault imposes an additional overhead of less than 1%. These overheads are acceptable in comparison to the alternative in which, currently, the effort expended on the calculation would be lost. Even in the ideal case where the triples would be restartable, the disruption caused by terminating and restarting the application would be significantly larger than the overhead of continuing the calculation in degraded mode.

We plan to deploy the infrastructure developed to other algorithms throughout NWChem. In particular, it is necessary to test the approach for algorithms where communication dominates more. The approach itself is not tied to NWChem, and we expect applications of it in other, most likely GA-based, codes as well.

**Acknowledgment.** This work was supported by the eXtreme Scale Computing Initiative at Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle. This work was done in part using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory, operated for the U.S. Department of Energy by Battelle under contract DE-AC05-76RL01830.

### References

(1) Jaguar Cray XT5 supercomputer. http://www.nccs.gov/jaguar/ (accessed Mar 31, 2010).

(2) The TOP500 report. http://www.top500.org/lists/2009/11 (accessed Mar 31, 2010).

(3) Amarasinghe, S.; Campbell, D.; Carlson, W.; Chien, A.; Dally, W.; Elnohazy, E.; Hall, M.; Harrison, R.; Harrod, W.; Hill, K.; Hiller, J.; Karp, S.; Koelbel, C.; Koester, D.; Kogge, P.; Levesque, J.; Reed, D.; Sarkar, V.; Schreiber, R.; Richards, M.; Scarpelli, A.; Shalf, J.; Snavely, A.; Sterling, T. Exascale software study: Software challenges in extreme scale systems, 2009. The office of Advanced Scientific Computing Research. http://www.er.doe.gov/ASCR/Research/CS/DARPAexascale-software(2009).pdf (accessed Mar 31, 2010).

(4) Valiev, M.; Bylaska, E.; Wang, D.; Kowalski, K.; Govind, N.; Straatsma, T.; van Dam, H.; Nieplocha, J.; Apra, E.;

Windus, T.; de Jong, W. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.

(5) Gordon, M.; Schmidt, M. Advances in electronic structure theory: GAMESS a decade later. In *Theory and applications of computational chemistry*; Dykstra, C., Frenking, G., Kim, K., Scuseria, G., Eds.; Elsevier: New York, 2005.

(6) Guest, M.; Bush, I.; van Dam, H.; Sherwood, P.; Thomas, J.; van Lenthe, J.; Havenith, R.; Kendrick, J. *Mol. Phys.* **2005**, *103*, 719–747.

(7) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2009.1; University College Cardiff Consultants Ltd.: Cardiff, United Kingdom, 2009.

(8) Karlström, G.; Lindh, R.; Malmqvist, P.-A.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neograd y, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(9) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

(10) Baker, J.; Wolinski, K.; Malagoli, M.; Kinghorn, D.; Wolinski, P.; Magyarfalvi, G.; Saebo, S.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2008**, *30*, 317–335.

(11) Janssen, C. L.; Nielsen, I. M. B. *Parallel computing in quantum chemistry*; CRC Press: Boca Raton, FL, 2008.

(12) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.

(13) *DALTON*, version 2.0; Chemistry Department at the University of Oslo: Oslo, Norway, 2005.

(14) Weber, V.; Challacombe, M. *J. Chem. Phys.* **2006**, *125*, 104110.

(15) Lischka, H.; Shepard, R.; Pitzer, R. M.; Shavitt, I.; Dallos, M.; T., M.; Szalay, P. G.; Seth, M.; Kedziora, G. S.; Yabushita, S.; Zhang, Z. *Phys. Chem. Chem. Phys.* **2001**, *3*, 664–673.

(16) Aprà, E.; Rendell, A. P.; Harrison, R. J.; Tipparaju, V.; deJong, W. A.; Xantheas, S. S. Liquid water: obtaining the right answer for the right reasons. In *SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; ACM: New York, 2009; pp 1−7.

(17) DeBardeleben, N.; Laros, J.; Daly, J.; Scott, S.; Engelmann, C.; Harrod, B. High-end computing resilience: Analysis of issues facing the HEC community and path-forward for research and development, 2010. Los Alamos National Laboratory Web Site. http://institute.lanl.gov/resilience/docs/HECResilienceWhitePaperJan2010final.pdf (accessed Mar 31, 2010).

(18) Daly, J. T. Application resilience for truculent systems. Presented at the 2009 Fault Tolerance Workshop for Extreme Scale Computing [Online], Albuquerque, NM, March 19−20, 2009. Teragrid Forum. http://www.teragridforum.org/mediawiki/images/8/80/Daly2009ws.pdf (accessed Mar 31, 2010).

(19) Nielsen, I. M. B.; Janssen, C. L.; Leininger, M. Scalable fault tolerant algorithms for linear-scaling coupled-cluster electronic structure methods, 2004. Sandia National Laboratories. http://prod.sandia.gov/techlib/access-control.cgi/2004/045462.pdf (accessed Sep 27, 2010).

(20) Nieplocha, J.; Harrison, R. J.; Littlefield, R. J. Global Arrays: A Portable "Shared-Memory" Programming Model for Distributed Memory Computers. In *Proc. Supercomputing'94*; IEEE CS Press: Washington, DC, 1994; pp 340−349.

(21) Harrison, R. *Int. J. Quantum Chem.* **1991**, *40*, 847–863.

(22) Lin, S.; Costello, D. J., Jr. *Error Control Coding: Fundamentals and Applications*, 2nd ed.; Prentice Hall: Englewood Cliffs, NJ, 2004.

(23) Rebaudengo, M.; Sonza Reorda, M.; Torchiano, M.; Violante, M. Soft-error detection through software fault-tolerance techniques. In *DFT '99, Proceedings of the 14th International Symposium on Defect and Fault-Tolerance in VLSI Systems, Albuquerque, NM, Nov 1−3, 1999*; IEEE Computer Society: Washington, DC, 2002; pp 210−218.

(24) Carlson, W. W.; Draper, J. M.; Culler, D. E.; Yelick, K.; Brooks, E.; Warren, K. Introduction to UPC and language specification, 1999. The George Washington University High Performance Computing Laboratory. http://www.gwu.edu/upc/publications/upctr.pdf (accessed July 28, 2010).

(25) Numrich, R. W.; Reid, J. *SIGPLAN Fortran Forum* **2005**, *24*, 4–17.

(26) Čìžek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266.

(27) Paldus, J.; Čìžek, J.; Shavitt, I. *Phys. Rev. A* **1972**, *5*, 50–67.

(28) Paldus, J.; Li, X. Z. Critical assessment of coupled cluster method in quantum chemistry. In *Advances in Chemical Physics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons, Inc.: New York, 1999; Vol 110, pp 1−175.

(29) Bartlett, R. J.; Musiał, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

(30) Kobayashi, R.; Rendell, A. P. *Chem. Phys. Lett.* **1997**, *265*, 1–11.

(31) Nieplocha, J.; Harrison, R. J.; Littlefield, R. J. *J. Supercomput.* **1996**, *10*, 169–189.

(32) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.

(33) Weikum, G.; Vossen, G. *Transactional information systems: Theory, algorithms, and the practice of concurrency control and recovery*; Morgan Kaufmann: San Francisco, CA, 2002.

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

(35) Chinook SuperComputer. http://www.emsl.pnl.gov/capabilities/computing/ (accessed Apr 26, 2010).

(36) Voltaire Technologies. http://www.voltaire.com/ (accessed Sep 27, 2010).

(37) Mellanox Technologies. http://www.mellanox.com/ (accessed Sep 27, 2010).

CT100439U

JCTC Journal of Chemical Theory and Computation

# Weak Molecular Interactions Studied with Parallel Implementations of the Local Pair Natural Orbital Coupled Pair and Coupled Cluster Methods

Dimitrios G. Liakos, Andreas Hansen, and Frank Neese*

*Lehrstuhl für Theoretische Chemie, Universität Bonn, Wegelerstrasse 12, D-53115 Bonn, Germany*

Received August 9, 2010

**Abstract:** A parallel implementation of the recently developed local pair natural orbital coupled electron pair approximation (LPNO-CEPA/*n*, *n* = Version 1, 2, or 3) and the corresponding LPNO coupled cluster method with single- and double excitations (LPNO−CCSD) is described. A detailed analysis alongside pseudocode is presented for the most important computational steps. The scaling with respect to the number of processors is reasonable and speedups of about 10 with 14 processors have been found in benchmark calculations (wall-clock time). The most important factor limiting the efficiency of the scaling with respect to the number of processors is probably the limited bandwidth of the presently prevailing multicore machines. The parallel LPNO methods were applied to study weak intermolecular interactions. Initially, the well-established S22 set of molecules was studied. The mean absolute error resulting from the use of the LPNO-CEPA/1 method relative to the most recent CCSD(T) reference data is found to be 0.24 kcal/mol. Thus, LPNO-CEPA/1 holds great promise for the efficient ab initio treatment of weak intermolecular interactions. In order to demonstrate the applicability of the methods to real systems, a two-dimensional potential energy surface for a trimer of 2,4-dihydroxy-3-acetyl-6-methyl acetophenone [$C_{11}H_{12}O_4$] (81 atoms, 1296 basis functions, 133 single points) has been calculated with the LPNO-CEPA/1 method. In this system, a clear distinction can be made between hydrogen bonding and $\pi-\pi$ interactions. The global minimum on the PES obtained from the calculations agrees excellently with the experimentally determined crystal structure. By contrast, popular density functional methods show no discernible minimum.

## Introduction

The ability to accurately and efficiently calculate electron correlation effects has been one of the major goals of electronic structure theory during the last decades. Density functional theory (DFT) represents the most popular family of methods in this respect.[1,2] Owing to their excellent price/performance ratio, DFT methods have become the standard tools for the calculation of molecular properties in most chemical applications. However, DFT methods do not represent a systematic hierarchy that converges to the exact solution of the Schrödinger equation, frequently contain empirical parameters, often are constructed with respect to arbitrarily chosen reference systems such as the inhomogeneous electron gas, and incorporate electron correlation effects in a somewhat ad hoc fashion. In particular, DFT methods do not incorporate dispersion effects and empirical corrections are required in order to obtain reasonable results for weakly bound systems (however, see recent efforts to incorporate the van der Waals interaction empirically[3−7] or by models based on physical reasoning[8−13]).

* Corresponding author e-mail: neese@thch.uni-bonn.de.

10.1021/ct100445s 
Published on Web 12/10/2010

An arguably more systematic route toward the incorporation of electron correlation effects is offered by wave function based ab initio methods, all of which take the Hartree−Fock (HF) determinant as the starting point. It has been realized that many body perturbation theory (MBPT) is often not robust enough for obtaining accurate results and only the lowest order correlation method, second order many body perturbation theory with Möller-Plesset partitioning (MP2) has survived as a widely used method owing to its comparatively small computational cost (for recent developments see refs 14−22). For accurate results, however, it is important to use iterative methods. The most popular family is based on the coupled cluster (CC) expansion that is size consistent, unitarily invariant, and converges reasonably quickly toward the full CI limit.[23−25] It is well-known that for obtaining accurate results, it is necessary to include connected triple excitations at least in a perturbative way, thus leading to the "gold" standard CCSD(T) method.[26−28] However, this method features seventh order scaling with respect to molecular size and is hence restricted to small molecules. An intermediate approach that lacks the rigor of CC theory but has been proven to lead to usefully accurate results (intermediate between CCSD and CCSD(T))[29] is the coupled electron pair approximation (CEPA[30−34]) which may be viewed as a simple size consistent modification of the configuration interaction methods with single- and double-excitations (CISD). CEPA is of slightly lower computational cost than CCSD but is usually more accurate.[29]

The most expensive steps in these methods involve multiple nested summations over the entire virtual orbital space. Many efforts have been done aiming in the development of low-order scaling correlation methods, like the ones from Werner and co-workers,[22,35−37] Head-Gordon and co-workers,[38−40] Ayala and Scuseria,[41,42] Carter and co-workers,[43,44] Auer and Nooijen,[45] and others. In this spirit, we have recently revived the method of pair natural orbitals (PNOs[30,31,33,34,46,47]) that greatly reduces the computational cost of these contractions and make the associated computational cost asymptotically linear scaling with respect to molecular size.[48,49] The PNO expansion can be viewed as a means of greatly compacting the information content contained in the virtual space thus leading to very compact correlated wave functions.[48] This is particularly important in combination with the large basis sets that are known to be required for obtaining accurate results in correlated ab initio calculations. In combination with localized internal orbitals and an electron pair screening procedure, the correlated singles- and doubles-wave function is expanded in a linear scaling number of wave function parameters. Our actual realization of the method in the framework of the CEPA and CCSD methods is not linearly scaling but rather retains some (small) fifth order steps.[48,49] Nevertheless, the methods have been proven to be efficient, reliable, of black box character and to recover typically 99.9% of the canonical correlation energy. In fact, some test calculations on molecules in the range of 40 to 80 atoms revealed that the actual time required for the correlation calculation was only two to four times larger than that taken by the proceeding Hartree−Fock calculations.[49] Thus, a very large range of chemical applications can be envisioned.

In this work, the applicability of the LPNO methods is further enhanced through the development of a parallel code. Furthermore, we demonstrate that the LPNO-CEPA/1 method provides accurate results for weak intermolecular interactions.

## Theory

The underlying theory of the LPNO-CEPA and LPNO−CCSD methods is described in detail elsewhere.[48,49] Here, we will only briefly describe the working equations in order to assist the understanding of the parallel program version.

**Working Equations.** Starting from the closed-shell CISD wave function and using the generator state formalism,[50] the CISD wave function is written as follows:

$$\Psi = \Psi_{\mathrm{HF}} + \sum_{ia} C_a^i \Psi_i^a + \sum_{i\le j}\sum_{ab} C_{ab}^{ij}\Psi_{ij}^{ab} \qquad (1)$$

with *i,j,k,l* referring to occupied orbitals and *a,b,c,d* to unoccupied ones. The residual for the double excitation amplitudes is as follows:

$$\sigma_{ab}^{ij} = \langle \tilde{\Psi}_{ij}^{ab} | H - E_0 - \Delta^{ij} | \Psi \rangle \qquad (2)$$

where $\Delta^{ij}$ is a method specific shift and $\tilde{\Psi}_{ij}^{ab}$denotes the contravariant configuration state function.[50] As derived in detailed elsewhere the working equation for the amplitudes in the PNO basis (the PNO basis is indicated by an overbar) becomes:

$$\begin{aligned}
\sigma_{\bar{a}^{ij}\bar{b}^{ij}}^{ij} &= K_{\bar{a}^{ij}\bar{b}^{ij}}^{ij} + K(\bar{\mathbf{C}}^{ij})_{\bar{a}^{ij}\bar{b}^{ij}} + \{\mathbf{d}^{ij\dagger}\mathbf{F}^{\mathbf{V}}\mathbf{d}^{ij}\bar{\mathbf{C}}^{ij} + \bar{\mathbf{C}}^{ij}\mathbf{d}^{ij\dagger}\mathbf{F}^{V}\mathbf{d}^{ij}\}_{\bar{a}^{ij}\bar{b}^{ij}} \\
&- \sum_{kl}(ik|jl)(\mathbf{S}^{ij,kl}\bar{\mathbf{C}}^{kl}\mathbf{S}^{ij,kl\dagger})_{\bar{a}^{ij}\bar{b}^{ij}} - \sum_{k}(\mathbf{F}_{ik}(\mathbf{S}^{ij,kl}\bar{\mathbf{C}}^{ik}\mathbf{S}^{ij,kl\dagger})_{\bar{a}^{ij}\bar{b}^{ij}} + \mathbf{F}_{ik}(\mathbf{S}^{ij,kj}\bar{\mathbf{C}}^{\mathrm{kj}}\mathbf{S}^{ij,kj\dagger})_{\bar{a}^{ij}\bar{b}^{ij}}) \\
&+ \sum_{k}\left\{\mathbf{S}^{ij,ik}(2\bar{\mathbf{C}}^{ik} - \bar{\mathbf{C}}^{ik\dagger})\left(\mathbf{K}^{kj} - \frac{1}{2}\mathbf{J}^{kj}\right)\mathbf{d}^{ij} + \mathbf{d}^{ij\dagger}\left(\mathbf{K}^{ik} - \frac{1}{2}\mathbf{J}^{ik}\right)(2\bar{\mathbf{C}}^{kj} - \bar{\mathbf{C}}^{kj\dagger})\mathbf{S}^{ij,ik\dagger}\right\}_{\bar{a}^{ij}\bar{b}^{ij}} \\
&- \sum_{k}\left\{\frac{1}{2}\mathbf{S}^{ij,ik}\bar{\mathbf{C}}^{ik\dagger}\mathbf{J}^{jk\dagger}\mathbf{d}^{ij} + \frac{1}{2}\mathbf{d}^{ij\dagger}\mathbf{J}^{ik}\bar{\mathbf{C}}^{kj\dagger}\mathbf{S}^{kj,ij} + \mathbf{d}^{ij\dagger}\mathbf{J}^{ik}\bar{\mathbf{C}}^{ik}\mathbf{S}^{ik,ij} + \mathbf{S}^{ij,kj}\bar{\mathbf{C}}^{kj}\mathbf{J}^{ik\dagger}\mathbf{d}^{ij}\right\}_{\bar{a}^{ij}\bar{b}^{ij}} \\
&+ C_{\bar{a}^{ij}}^{i}F_{\bar{b}^{ij}}^{j} + C_{\bar{b}^{ij}}^{j}F_{\bar{a}^{ij}}^{i} - \sum_{k}\{(jk|i\bar{a}^{ij})C_{\bar{b}^{ij}}^{k} + (ik|j\bar{b}^{ij})C_{\bar{a}^{ij}}^{k}\} \\
&+ \sum_{\bar{c}^{ij}}(i\bar{a}^{ij}|\bar{c}^{ij}\bar{b}^{ij})C_{\bar{c}^{ij}}^{j} + (i\bar{a}^{ij}|\bar{c}^{ij}\bar{b}^{ij})C_{\bar{c}^{ij}}^{j} - \Delta^{ij}\bar{\mathbf{C}}_{\bar{a}^{ij}\bar{b}^{ij}}^{ij}
\end{aligned} \qquad (3)$$

where $K_{ab}^{ij} = (ia|jb)$ and $J_{ab}^{ij} = (ij|ab)$are the usual exchange and Coulomb operators and two-electron integrals in round brackets are written in (11|22) notation. $S_{\bar{a}\bar{b}}^{ij} = \langle \bar{a}^{ij}|\bar{b}^{kl}\rangle = (\mathbf{d}^{ij\dagger}\mathbf{d}^{kl})_{\bar{a}\bar{b}}$ denotes the overlap between PNOs of different pairs. The matrices $\mathbf{d}^{ij}$are the transformation matrices from the canonical virtual basis to the PNO basis of pair $i,j$.[48]

Following Meyer,[47] $K(\bar{\mathbf{C}}^{ij})_{\bar{a}^{ij}\bar{b}^{ij}}$ denotes the 'external exchange' operator:

$$K(\bar{\mathbf{C}}^{ij})_{\bar{a}^{ij}\bar{b}^{ij}} = \sum_{\bar{a}^{ij}\bar{b}^{ij}} (\bar{a}^{ij}\bar{c}^{ij}|\bar{b}^{ij}\bar{d}^{ij})\bar{C}_{\bar{c}^{ij}\bar{d}^{ij}}^{ij} \tag{4}$$

The singles residual becomes:

$$\begin{aligned}\sigma_a^i &= F_a^i + \{\mathbf{F}^V\mathbf{C}^i\}_a - \sum_j F_{ij}C_a^j + \sum_{jkb}(2K_{jb}^{ik} - J_{jb}^{ik})C_{ba}^{kj} \\ &+ \sum_j \{(2\mathbf{K}^{ij} - \mathbf{J}^{ij})\mathbf{C}^j + \mathbf{F}^j(2\mathbf{C}^{ij\dagger} - \mathbf{C}^{ij})\}_a \\ &+ \sum_j \sum_{\bar{a}^{ij}} d_{a\bar{a}}^{ij} \sum_{\bar{b}^{ij}\bar{c}^{ij}} (2(i\bar{b}^{ij}|\bar{a}^{ij}\bar{c}^{ij}) - (i\bar{c}^{ij}|\bar{a}^{ij}\bar{b}^{ij}))C_{\bar{b}^{ij}\bar{c}^{ij}}^{ij} \\ &- \Delta^i C_a^i\end{aligned} \tag{5}$$

As explained in detail in refs,[51,52] the QCISD and CCSD methods can be implemented in a very similar fashion by introducing "dressed" integrals (to follow usual conventional we use cluster amplitudes $\mathbf{t}$ rather than CI coefficients $\mathbf{C}$ in these equations). The doubles residual becomes:

$$\begin{aligned}\sigma_{ab}^{ij} &= K_{ab}^{ij} + K(\tau^{ij})_{ab} + (\tilde{\tilde{\mathbf{F}}}^{\dagger}\tau^{ij} + \tau^{ij}\tilde{\tilde{\mathbf{F}}})_{ab} - \\ &\quad \sum_k (\tilde{\tilde{F}}_{jk}\mathbf{t}^{ki\dagger} + \tilde{\tilde{F}}_{ik}\mathbf{t}^{kj})_{ab} + \sum_{kl}{}''(ik|jl)\tau_{ab}^{kl} \\ &+ \sum_k \left((2\mathbf{t}^{ki\dagger} - \mathbf{t}^{ki})\left(\tilde{\mathbf{K}}^{jk\dagger} - \frac{1}{2}\tilde{\mathbf{J}}^{jk}\right) - \frac{1}{2}\mathbf{t}^{ki}\tilde{\mathbf{J}}^{jk} - \tilde{\mathbf{J}}^{jk\dagger}\mathbf{t}^{ki\dagger}\right)_{ab} \\ &+ \sum_k \left(\left(\tilde{\mathbf{K}}^{ik} - \frac{1}{2}\tilde{\mathbf{J}}^{ik\dagger}\right)(2\mathbf{t}^{kj} - \mathbf{t}^{kj\dagger}) - \frac{1}{2}\tilde{\mathbf{J}}^{ik\dagger}\mathbf{t}^{kj\dagger} - \mathbf{t}^{kj}\tilde{\mathbf{J}}^{ik}\right)_{ab} \\ &\quad - \sum_k ((jk|ia)t_b^k + (ik|jb)t_b^k) \\ &+ \sum_c ((ia|cb)t_c^j + (ib|ac)t_c^j) - \\ &\quad \left\{\sum_k ((\mathbf{K}^{ik}\mathbf{t}^j)_a t_b^k + (\mathbf{K}^{jk}\mathbf{t}^i)_b t_a^k) + (\mathbf{J}^{ik}\mathbf{t}^j)_b t_a^k + (\mathbf{J}^{jk}\mathbf{t}^i)_a t_b^k\right\}\end{aligned} \tag{6}$$

The dressed quantities are defined elsewhere.[49]

The singles sigma vector can be written in an analogous way as follows:

$$\begin{aligned}\sigma_a^i &= F_{ia} + \sum_b \tilde{F}_{ba}t_b^i - \sum_j \tilde{F}_{ij}t_a^j + G(\mathbf{t}_1)_{ia} + \sum_{jb}(2\mathbf{t}^{ji} - \mathbf{t}^{ji\dagger})\tilde{F}_{bj} \\ &- \sum_{kjb}(2(ik|jb) - (ij|kb))\tau_{ab}^{kj} \\ &+ \sum_{jbc}((2(ib|ac) - (ic|ab))\tau_{bc}^{ij} + (2(jb|ac) - (jc|ab))\tau_{cb}^{ij}) \\ &+ \left\{\sum_{jb}(\tilde{F}_{jb} - 2F_{jb})t_b^i t_a^j\right\}\end{aligned} \tag{7}$$

A complicating feature of LPNO-QCISD and LPNO-CCSD is that the dressed operators change in every iteration and can therefore not be precomputed over PNOs as in the case of LPNO-CEPA. How to overcome this problem is described in detail in ref 49.

**Parallel Implementation.** Our parallel implementation of the LPNO-CEPA and LPNO-CCSD methods, is based on standard Message Passing Interface[53] (MPI) libraries. The reason for this choice is portability of the code to both shared- and distributed memory machines. One important choice in the design of the parallel algorithm is the use of static or dynamic distribution of work across the processors. Dynamic distribution potentially offers the opportunity for better load balancing, especially when the number of processors is increasing. Our target platforms are clusters of personal computers (PCs) with a limited number of cores rather than massively parallel computers. Given that load-balancing problems do not appear to be overwhelming (vide infra), we have chosen the static distribution model, which potentially also offers the significant advantage of data distribution across local hard drives.

There are three computationally expensive parts of a LPNO-CEPA/CCSD implementation: (a) the canonical integral transformation, (b) the transformation of various classes of two-electron electron repulsion integrals into the PNO basis, and (c) the calculation of the sigma vector. The parallelization of these steps will be described in detail below. The algorithm discussed below has been implemented into the ORCA electronic structure package and is already publically available.[54]

**Canonical Integrals.** The canonical integral transformations in the LPNO methods are based on the Resolution of the Identity (RI) approximation.[55] In the first step of the transformation, the 3-index repulsion integrals over basis functions are calculated, transformed to the molecular orbital basis, and stored on disk. However, we prefer to avoid the generation, storage and resorting of the largest class of such integrals, $(ab|\tilde{K})$ ($\tilde{K}$ refers to an orthogonalized auxiliary basis function in the Coulomb metric[49]). In the second step, the necessary four index integrals are formed from the prestored three index integrals. The first step is computationally insignificant and will therefore not be further described.

In the second half transformation, the integrals that are required for a LPNO calculation are generated: (a) all internal $(ik|jl)$, b) one external $(ik|ja)$, $(ij|ka)$, and c) two external $(ia|jb)$, $(ij|ab)$. By far the largest set of integrals is the two external integrals and thus special attention should be given to their generation. Given that the set of one-external RI integrals $(ia|\tilde{K})$ is rather small, they can conveniently stored on disk. The generation of $(ia|jb)$ then proceeds efficiently (as in the case of RI-MP2[56]) by efficient matrix multiplications. We store matrices ordered by the internal index to this end ($X_{aK}^i = (ia|\tilde{K})$).

The generation of the Coulomb integrals, $(ij|ab)$, is more difficult as a convenient matrix driven strategy is less obvious. Since the generation of the three-index RI integrals over atomic orbitals, $(\mu\nu|K)$, is computationally inexpensive compared to the remaining steps, we have developed an integral direct algorithm for the generation of $(ij|ab)$ with the intentions to: (a) minimize input/output (I/O) overheads, (b) avoid lengthy integral sorts, (c) use efficient BLAS level 3 matrix multiplications to the largest possible extent within

```
Read in all-internal RI integrals
Construct batches according to number of processors and available memory
For batches of indices ab distributed across processors
    For K in auxiliary basis functions
          for μ in atomic orbitals
                for ν in atomic orbitals (μ≤ν)
                      Prescreen for negligible integrals
                      Calculate Y^K(μ,ν)=(μν|K).
                End ν
          End μ
          Transform X(K,ab)=(K|ab)=(c_L^T Y^K c_R)_ab
            (using batch specific truncated MO coefficient matrices c)
    End K.
    Orthogonalize the integrals Z= V^-1/2 X.
    Generate subset of target integrals J^ij(a,b)=(ij|K)*Z(K,ab)
                                    (using matrix multiplications)
    Gather the local results
    Store the integrals.
End batches
```

***Figure 1.*** Pseudo code for the parallel generation of (*ij*|*ab*) within the RI approximation as implemented in ORCA.

a given amount of core memory, and (d) achieve good parallel scalability.

This strategy implies a batch algorithm that is structured as shown in Figure 1. It is assumed that the all internal RI integrals, (*ij*|*K*), fit into memory and this assumption has so far never created a bottleneck in actual calculations. In each batch, a loop over all internal pairs is performed and as many external pairs as fit in memory are treated simultaneously. In the parallel algorithm, the work is distributed over external pairs thus leading to fine-grained parallelization. Inside each batch, the RI integral generation is driven by an outer loop over auxiliary functions *K*. For each *K* within a given angular momentum shell, the entire matrix of atomic integrals $Y^K_{\mu\nu} = (\mu\nu|K)$ is generated. From these integrals, BLAS level 3 matrix multiplications are used to generated as many (*ab*|*K*) as memory permits. Truncated MO coefficient matrices are set up to this end. These integrals are orthogonalized through a second matrix multiplication with the $\mathbf{V}^{-1/2}$ matrix ($V_{KL} = (K|L)$ is the Coulomb metric) or an equivalent thereof. The orthogonalized integrals are contracted in the third step with $(ij|\tilde{K})$ in order to generate a subset of the target (*ij*|*ab*) integrals. For storage of these and similar integrals, a special data structure is employed (referred to as "Matrix Container" in ORCA), where integrals are stored as matrices ordered by the internal pair labels ($\mathbf{J}^{ij}$,$\mathbf{K}^{ij}$) as required by the matrix formulation of the CCSD and CEPA equations (eqs 3, 5, 6, and 7). In the parallel algorithm, a limited amount of overhead arises from gathering the partial $\mathbf{J}^{ij}$ operators formed at different processors before writing them to disk. This procedure also necessarily synchronizes the processes. The rest of the algorithm does not require communication and proceeds as efficiently as in the single processor case since the size of the matrix multiplications are independent of the number of processors. The disk storage of the integrals is organized in a "per node" fashion, where the integrals are stored only for each node and not for each processor or core. This implies that if all processors are located on the same node, the integrals are stored only once and are available to all processors. As long as the communication overhead remains small and the generation of the RI-integrals over atomic orbitals is not rate limiting, the algorithm is expected to scale nearly linearly with the number of processes.

***Table 1.*** Wall-Clock Timings and Speedups for the Parallel Creation of the (ij|ab) Integrals for **1**[a]

| | processors | | | | | | |
|---|---|---|---|---|---|---|---|
| direct_Jij | 1 | 2 | 4 | 6 | 8 | 10 | 14 |
| time (sec) | 71008 | 33785 | 19269 | 13149 | 10611 | 8375 | 7424 |
| speedup | 1 | 2.1 | 3.7 | 5.4 | 6.7 | 8.5 | 9.6 |

[a] Basis Set: def2-TZVP[58,59] (1130 basis functions). Auxilliary Basis Set: def2-TZVP/C (3122 basis functions).

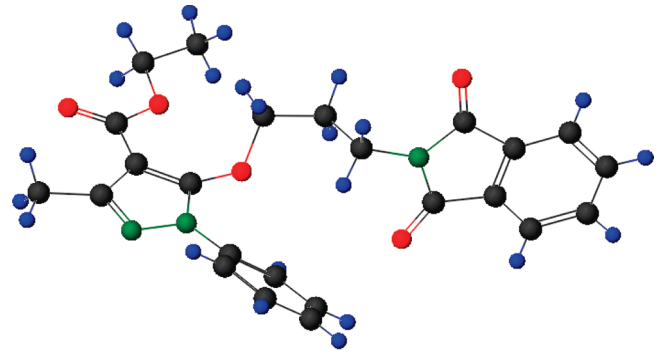


***Figure 2.*** The structure of **1** = 5-[3-(1,3-dioxo-1,3-dihydro-isoindol-2-yl)-propoxy]-3-methyl-1-phenyl-1H-py pyrazole-4-carboxylic acid ethyl ester ($C_{24}H_{23}N_3O_5$).

In Table 1, we present timings for the transformation. For all timings presented in this work, we used the molecule 5-[3-(1,3-dioxo-1,3-dihydro-isoindol-2-yl)-propoxy]-3-methyl-1-phenyl-1H-py pyrazole-4-carboxylic acid ethyl ester ($C_{24}H_{23}N_3O_5 \equiv$ **1**, see Figure 2).[57] The def2-TZVP[58,59] basis set is used (1130 contracted basis functions), together with the corresponding def2-TZVP/C auxiliary basis (3122 auxiliary basis functions).

It is observed that the scaling with the number of processes is, unfortunately, not perfectly linear. We believe that the reason for this is that the calculations were performed on a machine with four AMD 4 Quad-Core AMD Opteron 8354 processors per node. For these machines, it is known that memory bandwidth becomes rate limiting for compute and memory intensive operations.[60,61] However, a speedup of 9.6 is still observed with 14 processors. For parallel machines with better memory bandwidth better scalability is expected.

```
For K=MyID  to K = Naux, K+=NumProcs
    For μ in basis functions
          For ν in basis functions (μ ≤ ν)
                Prescreen for negligible integrals
                Calculate (μν|K)
          End ν
    End μ
    Transform (μν|K) to (ij|K), (ia|K) and (ab|K)
    For ip(ij) in internal pairs
          Create and store three index integrals over the PNOs of pair ip
    End ip
End K
Collect local parts of integrals.
For ip=MyID to ip < Internal pairs. ip += NumProcs
    Copy the relevant part of the metric matrix for pair ip
    Create local orthogonalization matrix for pair ip
    Use this matrix to transform the PNO repulsion integrals for pair ip.
    Construct the final 4-index repulsion integrals for pair ip.
End ip
Gather local parts
```

***Figure 3.*** Pseudo code for the parallel local transformation of the repulsion integrals to the PNO basis.

**PNO Transformation.** The second computationally expensive step is the generation of electron−electron repulsion integrals over PNOs. Since the number of PNOs is typically much larger than the number of virtual orbitals, the transformation must be carefully arranged in order to avoid computational bottlenecks (Figure 3). As discussed in the original work,[48] the RI approximation is used to generate the integrals over PNOs. The procedure is divided in two steps. In the first step, the RI integrals over AOs are generated ($(\mu\nu|K)$) with the auxiliary index running slow, as described above for the canonical integrals. From these integrals, we first produce the canonical integrals $(ij|K)$, $(ia|K)$, and $(ab|K)$ for a given set of K's that belong to a given angular momentum shell inside the auxiliary basis set. This transformation is done using BLAS level 3 operations. Hence, sparsity and integral prescreening is used in the integral generation but not in the initial transformation. A loop over pairs $P \equiv (ij)$ is next performed. Inside this loop, we first generate and store the integrals $(ij|K)$, $(i\bar{a}ij|K)$, $(j\bar{a}ij|K)$, and $(\bar{a}ij\bar{b}ij|K)$. Since the number of these integrals is linear scaling, the storage creates no bottlenecks despite the fact that there may be tens of thousands of PNOs in the calculation. It should be noted that the algorithm contains pair specific local fitting domains such that only those pairs that have the auxiliary index $K$ inside their domain are treated. As explained in detail in reference,[48] the domain construction is organized using a highly conservative threshold ($T_{\mathrm{CutMKN}} = 10^{-3}$) such that the error introduced is negligible. The loop ends by synchronization in which all integrals are communicated to all processors.

The last step of the PNO integral transformation algorithm is driven by a loop over internal electron pairs. For each pair, the local Coulomb metric is created and the pair specific three index repulsion integrals created in the previous step are orthogonalized to give three-index repulsion integrals over orthogonalized auxiliary functions $\tilde{K}$. This enables the calculation of the target integrals $(i\bar{a}^{ij}|j\bar{b}^{ij})$, $(ij|\bar{a}^{ij}\bar{b}^{ij})$, $(i\bar{a}^{ij}|\bar{b}^{ij}\bar{c}^{ij})$, $(j\bar{a}^{ij}|\bar{b}^{ij}\bar{c}^{ij})$, and $(\bar{a}^{ij}\bar{b}^{ij}|\bar{c}^{ij}\bar{d}^{ij})$ through efficient dense matrix multiplications. These integrals are stored on disk. This step of the algorithm is linear scaling and is, once more, parallelized in a round robin fashion by distributing the loop over pairs over all processors.

***Table 2.*** Wall-Clock Timings and Speedups for the Integral Parallel Transformation to the PNO Basis during a LPNO−CCSD Calculation on **1**[a]

| PNO transformation | processors | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 4 | 6 | 8 | 10 | 14 |
| time (sec) | 39494 | 25144 | 11691 | 8292 | 6990 | 7658 | 4821 |
| speedup | 1 | 1.6 | 3.4 | 4.8 | 5.7 | 5.2 | 8.2 |

[a] Basis Set: def2-TZVP[58,59] (1130 basis functions). Auxilliary Basis Set: def2-TZVP/C (3122 basis functions).

The final step of the PNO integrals transformation consists of the calculation of overlap matrices between the PNOs of different pairs. These overlap matrices are stored in packed form on disk. Only a linear number of overlap integrals is nonvanishing such that no storage bottlenecks arise. The overlap integrals are calculated through efficient matrix multiplications.

In Table 2, we present the time and speed-ups for the PNO transformation. Once more, it is noted that the algorithm does not scale perfectly linearly with the number of processors. Again, we believe that this is due to limited memory bandwidth of the machines used. However, a speedup of 8.2 is still observed for 14 processors.

**Sigma Vector Construction.** The third major part of the LPNO-CEPA/CCSD methods is the construction of the sigma vector. Below, we will comment on how the computationally most significant terms were implemented in the parallel case.

**Singles Fock.** The construction of the singles Fock matrix $G(t_1)$ is presently the most time-consuming step in the sigma-vector construction during a LPNO−CCSD calculation. As discussed in ref 49, this step is not necessary for LPNO-CEPA or LPNO-QCISD which hence do have a timing advantage over LPNO−CCSD. The advantages are most pronounced for highly polarized basis sets where integral evaluation becomes computationally expensive. The parallelization of this step is, however, straightforward and follows the techniques that are available throughout the ORCA program. The load balancing obtained during Fock-matrix construction is excellent and the direct computation of the singles-Fock matrix scales nearly linearly with the number of processors up to 14 processors. The only communication

required is the gathering of the partial Fock-type matrices after the integral evaluation has been completed. It is to be expected that this step will significantly improve through the use of the RIJCOSX[62] approximation.

**Doubles−Doubles Interaction with No External Labels.** The computationally next most demanding term is the disjoint doubles−doubles interaction that involves the all-internal exchange integrals (*ik*|*jl*). These terms are written in canonical form as follows:[48]

$$\sum_{kl} K_{kl}^{ij} t_{ab}^{kl} - \sum_{k} \{F_{jk} t_{ab}^{ik} + F_{ik} t_{ab}^{kj}\}$$

and in the LPNO case as follows:

$$\sum_{kl} (ik|jl)(\mathbf{S}^{ij,kl}\bar{\mathbf{t}}^{kl}\mathbf{S}^{ij,kl\dagger})_{\bar{a}^{ij}\bar{b}^{ij}} - \sum_{k} (F_{jk}(\mathbf{S}^{ij,ik}\bar{\mathbf{t}}^{ik}\mathbf{S}^{ij,ik\dagger})_{\bar{a}^{ij}\bar{b}^{ij}} + F_{ik}(\mathbf{S}^{ij,kj}\bar{\mathbf{t}}^{kj}\mathbf{S}^{ij,kj\dagger})_{\bar{a}^{ij}\bar{b}^{ij}})$$

The algorithm commences with a double loop over *k* and *l* followed by reading a matrix of integrals, $K_{ij}^{kl} = (ki|lj)$. This is followed by a loop over pairs P with internal indices *i* and *j*. Pair−pair interactions with negligible values of $|(ki|lj)|$ are dropped. The remaining significant contributions are calculated by transforming the amplitudes of pair *k*,*l* into the PNO basis of pair *i*,*j* and update of the sigma-vector. Parallelization takes place by distributing the pair loop over processors. There is no communication between processors required in this step as the local sigma vector components are gathered at the end of each coupled pair/coupled cluster iteration. Thus, the only issue to be addressed is load balancing that is once more done in a round-robin fashion. Thus, this step should scale linearly with the number of processors as the load balancing will tend to become perfect for larger molecules.

**Doubles-Singles Interaction with One External Label.** Quite unexpectedly, a step that becomes computationally significant in the LPNO methods is the interaction of the double excitations with singles through integrals with one-external label. The origin of this is that the singles are left in canonical form. The sigma-vector contribution is written as follows:

$$\sigma_a^i = \sum_j \{(2\mathbf{K}^{ij} - \mathbf{J}^{ij})\mathbf{t}^j + \mathbf{F}^j(2\mathbf{t}^{ij\dagger} - \mathbf{t}^{ij})\}$$

The complication arises from the fact that the doubles amplitudes have to be in the canonical basis for the calculation of this term. Hence, the back transformation from the PNO to the canonical basis must be performed inside a pair loop. It is to be expected that the calculation of this term can be significantly improved through prescreening or a more judicious choice if orbitals to expand the single excitations in.

Parallelization is achieved by dividing the pair loop over processors in a round-robin fashion. During the loop over pairs, there is no communication needed among the processors and the parallel overhead comes just in the form of gathering the individual parts of the singles sigma vectors, which is computationally insignificant. Hence, this part of the calculation should scale linearly with the number of processors.

***Table 3.*** Wall-Clock Timings and Speedups for the Creation of the Sigma Vector in a LPNO−CCSD Calculation on **1**[a]

| sigma vector | processors | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 4 | 6 | 8 | 10 | 14 |
| time (sec) | 188857 | 99326 | 53619 | 34203 | 24420 | 22029 | 15024 |
| speedup | 1 | 1.9 | 3.5 | 5.5 | 7.7 | 8.6 | 12.6 |

[a] Basis Set: def2-TZVP[58,59] (1130 basis functions). Auxilliary Basis Set: def2-TZVP/C (3122 basis functions).

**Doubles-Doubles Interaction with Two External Labels.** As discussed at lengths in refs 48 and 49 this term can be written as follows:

$$\sum_k \left\{ \begin{array}{c} \mathbf{S}^{ij,ik}(2\bar{\mathbf{t}}^{ik} - \bar{\mathbf{t}}^{ik\dagger})\left(\mathbf{K}^{kj} - \frac{1}{2}\mathbf{J}^{kj}\right)\mathbf{d}^{ij} + \\ \mathbf{d}^{ij\dagger}\left(\mathbf{K}^{ik} - \frac{1}{2}\mathbf{J}^{ik}\right)(2\bar{\mathbf{t}}^{kj} - \bar{\mathbf{t}}^{kj\dagger})\mathbf{S}^{ij,ik\dagger} \end{array} \right\}_{\bar{a}^{ij}\bar{b}^{ij}} -$$

$$\sum_k \left\{ \begin{array}{c} \frac{1}{2}\mathbf{S}^{ij,ik}\bar{\mathbf{t}}^{ik\dagger}\mathbf{J}^{jk\dagger}\mathbf{d}^{ij} + \frac{1}{2}\mathbf{d}^{ij\dagger}\mathbf{J}^{ik}\bar{\mathbf{t}}^{kj\dagger}\mathbf{S}^{kj,ij} + \\ \mathbf{d}^{ij\dagger}\mathbf{J}^{ik}\bar{\mathbf{t}}^{ik}\mathbf{S}^{ik,ij} + \mathbf{S}^{ij,kj}\bar{\mathbf{t}}^{kj}\mathbf{J}^{ik\dagger}\mathbf{d}^{ij} \end{array} \right\}_{\bar{a}^{ij}\bar{b}^{ij}}$$

As explained elsewhere,[49] this term greatly profits from precalculating the contractions of the canonical **J***ij* and **K***ij* operators with the relevant PNO transformation matrices **d**. During the iterations, these partial integrals are read and contracted with the PNO amplitudes of the interacting pair in order to arrive at the final sigma-vector contributions.

In order to parallelize this contribution, the first step consists of the parallelization of the creation of the partial integral file. This is achieved by treating only a subset of internal pairs on each processor. This process creates a local partial pair interaction file. Hence, during sigma-vector construction each processor only processes this partial file which means that no further explicit parallelization is necessary.

**Parallel Speedup of Sigma-Vector Contributions.** In Table 3 and Figure 4, the absolute times and speedups for the creation of the sigma vector for the test molecule in Figure 2 are shown. It is observed that up to 14 processors the program scales very well, as expected from the analysis presented above. In Table 4 the wall-clock times for the complete LPNO-CCSD calculation are analyzed. An overall speedup of 9.4 is observed for 14 processors. While this is less than the ideal result, we have argued above that the parallel efficiency is presently mainly limited by memory bandwidth of the machines used and hence may well improve for future generation hardware. We have nevertheless chosen to use the present machines as they represent typical hardware for computational chemistry applications. The main result is that the wall clock times required for LPNO-CCSD calculations on standard hardware can be reduced by an order of magnitude through parallelization.

## Application to Weak Interactions

Noncovalent interactions play a key role in modern chemical research, for example in supramolecular chemistry. Hence,

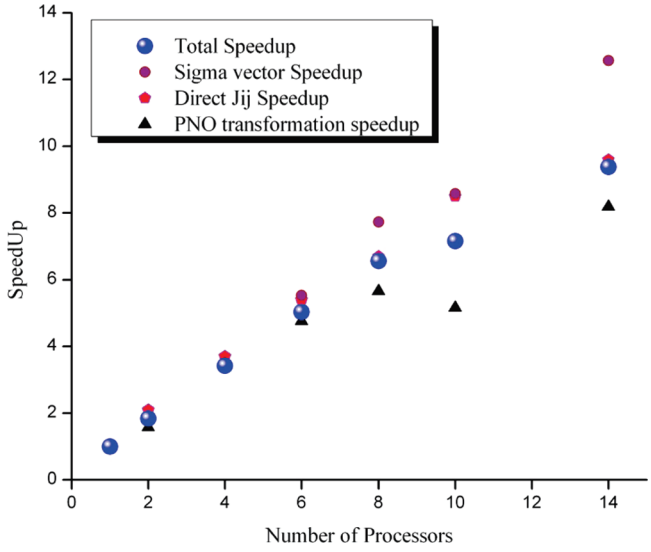



***Figure 4.*** Plot of the speedups for a LPNO-CCSD calculation on **1** versus the number of processors used. A computer with four Quad-Core AMD Opteron 8354 Processors per node was used for the calculations. The size of the cache memory was 512 Kbytes at the L2 level and 2 Mbytes at the L3 level. Overall, 120 GBytes of memory were available for the whole node and the program used 1.5 GBytes per processor.

***Table 4.*** Wall-Clock Timings and Speedups for a Complete LPNO−CCSD Calculation on **1**[a]

| total LPNO−CCSD | processors | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 4 | 6 | 8 | 10 | 14 |
| time (sec) | 336789 | 183084 | 98674 | 66973 | 51393 | 47099 | 35927 |
| speedup | 1 | 1.8 | 3.4 | 5.0 | 6.6 | 7.2 | 9.4 |

[a] Basis Set: def2-TZVP[58,59] (1130 basis functions). Auxilliary Basis Set: def2-TZVP/C (3122 basis functions).

this is also an active area of theoretical research.[63−67] The description of weak intermolecular forces presents a challenge for the theoretical chemistry since highly accurate methods such as CCSD(T) are computationally too expensive for routine use and computationally cheaper methods (mostly based on DFT) are not accurate enough.[68−71] (However, see recent results of Grimme and co-workers[6,56,72,73]). As will be shown below, the LPNO methods offer a computationally tractable route to this problem without introducing any element of empiricism.

**S22 Benchmark Calculations.** In order to evaluate the accuracy of the LPNO based methods for intermolecular interactions, the defacto standard S22 set of molecules was treated.[74] After the original publication in 2006, two recent articles have been published[75,76] that improve upon the reference values. In this work, the results of Takatani et al.[76] were used as reference. The geometries provided in the original publications were used throughout. We have also followed the same extrapolation scheme for the estimation of the complete basis set (CBS) limit. The complete basis set (CBS) MP2 correlation energy is estimated as follows:[77]

$$E_{\mathrm{MP2}}^{(\infty)} = \frac{X^3 E_{\mathrm{MP2}}^{(X)} - Y^3 E_{\mathrm{MP2}}^{(Y)}}{X^3 - Y^3} \qquad (8)$$

***Table 5.*** Calculated Reaction Energies for the S22 Set Using the LPNO-CEPA/1 Method[a]

| complex | ref 76 | calculated energy | error |
|---|---|---|---|
| $(NH_3)_2$ | −3.17 | −3.03 | −0.14 |
| $(H_2O)_2$ | −5.02 | −4.93 | −0.09 |
| formic acid dimer | −18.8 | −18.28 | −0.52 |
| formamide dimer | −16.12 | −15.61 | −0.51 |
| uracil dimer ($C_{2\ h}$) | −20.69 | −19.94 | −0.75 |
| 2-pyridoxine.. 0.2-amino-pyridine | −17 | −16.24 | −0.76 |
| adenine...thymine (WC) | −16.74 | −15.95 | −0.79 |
| $(CH_4)_2$ | −0.53 | −0.49 | −0.04 |
| $(C_2H_4)_2$ | −1.5 | −1.45 | −0.05 |
| $C_6H_6...CH_4$ | −1.45 | −1.53 | 0.08 |
| $C_6H_6...C_6H_6$(stacked) | −2.62 | −2.59 | −0.03 |
| pyrazine dimer | −4.2 | −4.01 | −0.19 |
| uracil dimer ($C_2$) | −9.74 | −9.45 | −0.29 |
| indole...benzene (stacked) | −4.59 | −4.62 | 0.03 |
| adenine...thymine (stacked) | −11.66 | −11.14 | −0.52 |
| $C_2H_4...C_2H_2$ | −1.51 | −1.54 | 0.03 |
| $C_6H_6...H_2O$ | −3.29 | −3.27 | −0.02 |
| $C_6H_6...NH_3$ | −2.32 | −2.37 | 0.05 |
| $C_6H_6...HCN$ | −4.55 | −4.63 | 0.08 |
| $C_6H_6...C_6H_6$ (T-Shape) | −2.71 | −2.84 | 0.13 |
| indole...benzene(T-shape) | −5.62 | −5.71 | 0.09 |
| phenol dimer | −7.09 | −6.89 | −0.20 |
| mean absolute error | | | 0.24 |
| max error | | | 0.79 |

[a] All values in kcal/mol.

Here $X$ and $Y$ are ($Y > X$) the two cardinal numbers of the two basis sets used for the two-point extrapolation and $E_{\mathrm{MP2}}^{(X)}$,$E_{\mathrm{MP2}}^{(Y)}$ are the MP2 correlation energies calculated with the two basis sets. The total energy in the CBS limit is then estimated as follows:

$$E_{\mathrm{total}}^{(\mathrm{CBS})} \approx E_{\mathrm{SCF}}^{Y} + E_{\mathrm{corr}}^{\mathrm{LPNO};X} + E_{\mathrm{corr}}^{(\mathrm{MP2};\infty)} - E_{\mathrm{corr}}^{(\mathrm{MP2};X)} \qquad (9)$$

Here $E_{\mathrm{corr}}^{(\mathrm{LPNO};X)}$ is the correlation energy calculated with LPNO-CEPA/1 within the smaller basis set and $E_{\mathrm{SCF}}^{Y}$ the SCF energy calculated with the larger basis set. Thus, it is assumed that the SCF energy with the larger basis is close enough to the basis set limit and that the difference between the MP2 basis set limit and the MP2 energy calculated with the small basis set is a reliable estimate of the difference in correlation energies that would be obtained with the high-level correlation method (in this case LPNO-CEPA/1). The accuracy of this scheme will be investigated in detail elsewhere.

For all molecules the aug-cc-pVTZ/aug-cc-pVQZ bases[78] together with their corresponding auxiliary basis sets for the RI approximation were employed. No counterpoise correction was employed as this is inconsistent with the notion of CBS extrapolation.

In Table 5, the calculated reaction energies are presented together with the deviations from the reference values and the calculated mean absolute error (MAE). It is observed that relative to the reference CCSD(T) values, the MAE of the LPNO-CEPA/1 method is only 0.24 kcal/mol and the maximum error is 0.79 kcal/mol. These results imply that the LPNO-CEPA/1 method is an accurate and computationally efficient method for the treatment of weak, noncovalent interactions. Importantly, this method can be applied to much larger systems than the canonical CEPA, CCSD, or CCSD(T) methods.

***Table 6.*** Mean Absolute Errors for the S22 Set Using Various Electronic Structure Methods

| method | type | mean absolute error (kcal/mol) |
|---|---|---|
| LPNO-CEPA/1 | | 0.24 |
| PW86PBE-XDM(BR) | GGA-XDM | 0.34 |
| B2PLYP-D | double hybrid-D | 0.43 |
| M06−2X/MG3S | hybrid meta-GGA | 0.43 |
| B97-D | GGA-D | 0.46 |
| B3LYP-D | hybrid GGA-D | 0.52 |
| PW86PBE-XDM(XX) | GGA-XDM | 0.72 |
| B2PLYP | double hybrid | 1.79 |
| B3LYP | hybrid GGA | 3.77 |
| BP86 | GGA | 4.14 |

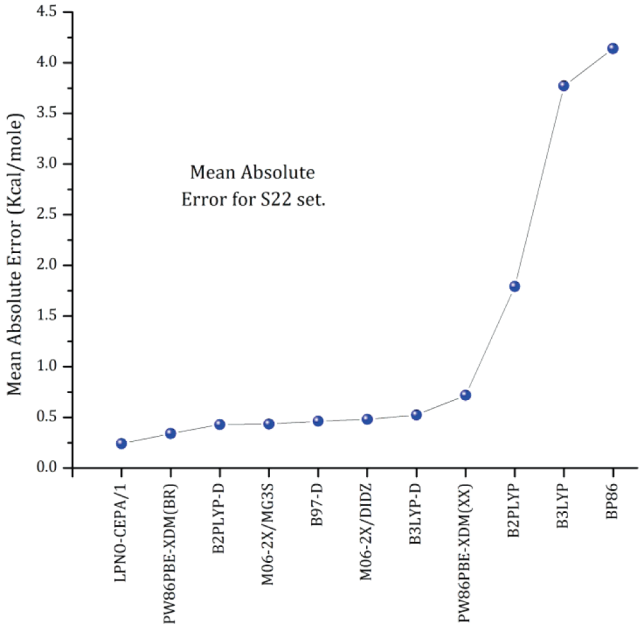



***Figure 5.*** Mean Absolute Error for the 22 reactions of the S22 set.

In order to compare our results with the ones obtained by present day DFT functionals, representative results are collected in Table 6 and are graphically shown in Figure 5. While the collection of functionals represent only a small fraction of the available literature, they do represent the presently prevailing classes of functionals (Zhao et al.[13] present a comparison for a more extensive set of functionals). Specifically, we used BP86[79,80] (values taken from reference[53]) as an example of a GGA functional, B3LYP[81−83] as a typical hybrid functional (values taken from Prof. Grimmes Web site[84]) and B2PLYP[85] as a double hybrid functional (values taken from Prof. Grimmes Web site[84]). These classes of functional were then considered together with the semiempirical van der Waals correction proposed by Grimme (B97-D,[86] B3LYP-D,[3,4] and B2PLYP-D[3,4] functionals as dispersion corrected GGA, hybrid and doubly hybrid functionals respectively. M06−2X/MG3S47[71] and M06−2X/DIDZ47[71] (values taken from ref 71) were used as typical hybrid meta-GGA functionals. Finally, the recently proposed PW86PBE-XDM(BR)[8] and PW86PBE-XDM(XX)[8] functionals that are specifically tailored to the calculation of noncovalent interactions were considered. It is pleasing to observe that

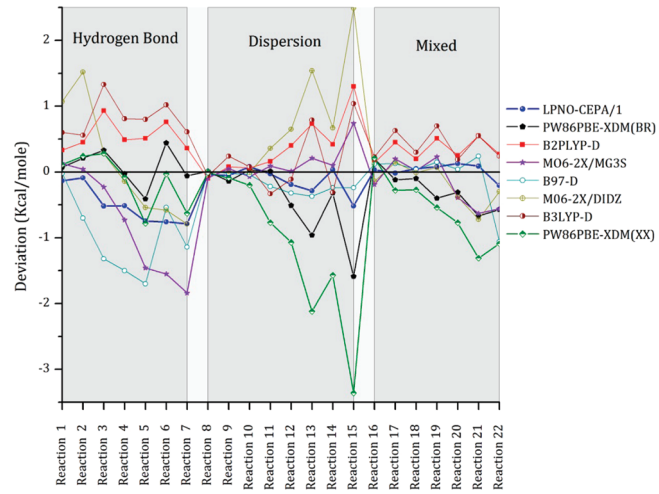



***Figure 6.*** Deviations from the reference values for the S22 set obtained by different computational methods as described in the text.

of all methods tested, the LPNO-CEPA/1 methods leads to the smallest mean absolute error. Thus it even outperforms functionals that are specifically designed for the calculation of weak interactions. That LPNO-CEPA/1 is also highly successful for many other molecular properties[48] and hence represents an accurate and efficient model chemistry.

In order to analyze these results more closely, Figure 6 presents a plot collecting the deviations from the reference values for different methods. The plot has three parts describing reactions concerning complexes with mainly hydrogen bonding interactions, complexes where dispersion forces are dominant and finally complexes with both types of interactions. It is concluded that none of the tested functionals can treat both kind of interactions with consistent accuracy. For example PW86PBE-XBM(BR), which presented the lowest mean absolute error of all functionals, seems to describe hydrogen bonds accurate but shows errors of up to 1.6 kcal/mol for the dispersion dominated cases. Similarly, B97-D describes dispersion very well but is less accurate for hydrogen bonds. The same is true for LPNO-CEPA/1, where however, the largest absolute error is still smaller than 1 kcal/mol.

**Application to a Trimer System.** In order to demonstrate the application of LPNO-CEPA/1 to a larger "real-life" system, 2,4-dihydroxy-3-acetyl-6-methyl acetophenone [$C_{11}H_{12}O_4$] (≡ **2**, see Figure 7) was treated. This system has been recently synthesized and characterized by spectroscopic as well as X-ray diffraction studies.[87] It was found that in the crystal structure both hydrogen bonds and $\pi-\pi$ stacking interactions are crucial for the formation of a two-dimensional supramolecular assembly (Figure 7).

The crystal structure of **2** shows that there are discrete trimers formed and it is evident that the $\pi$-stacking and hydrogen bonding interactions can be nicely differentiated. Monomer A interacts with monomer B through two hydrogen bonds while the interaction between monomers B and C occurs through $\pi-\pi$ stacking interactions.

Since the two interactions are fairly well separated, we have conjectured that only two geometrical parameters are

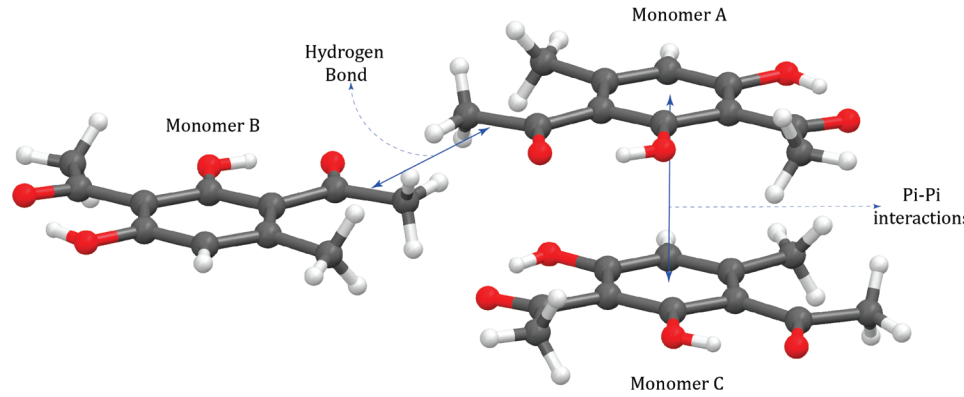



***Figure 7.*** The molcular structure of **2** = (2−4-dihydroxy-3-acetyl-6-methyl acetophenone)$_3$.

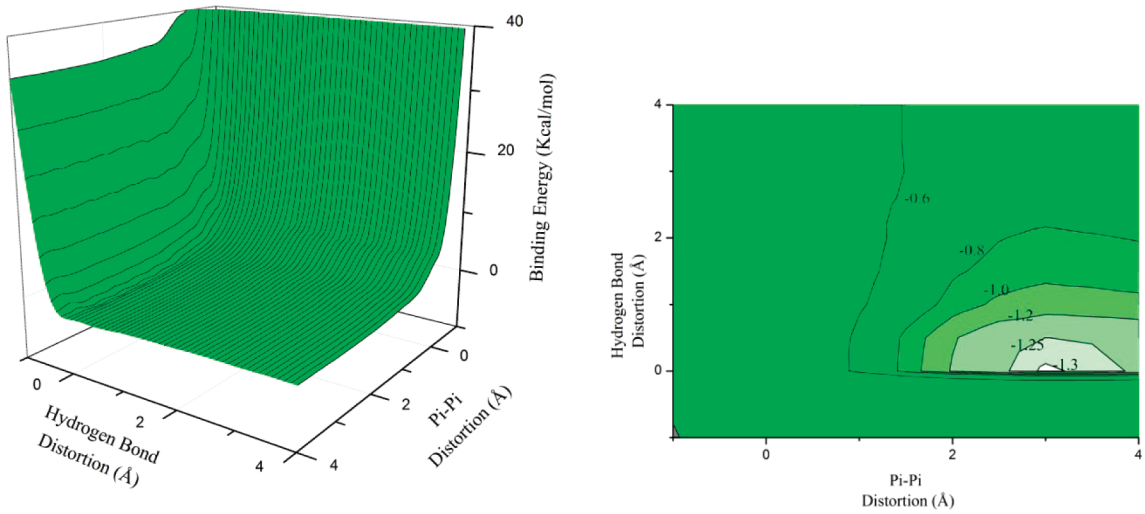



***Figure 8.*** The PES of **2** calculated at the BP86/def2-TZVP level of theory. Left: 3-dimensional surface plot, Right: Contour plot.

of key importance. The hydrogen bond contact interactions and the $\pi-\pi$ stacking distance. The total energy was scanned along these two coordinates (see Figure 5) while keeping the remaining internal structures of the building blocks fixed. Eleven steps between −1 and 4 Å (relative to the experimental structure) were performed for each coordinate. Additionally close to the potential minimum, a finer mesh of 0.2 Å was used. Overall, 133 points were calculated. For comparison, both BP86 and LPNO-CEPA/1 calculations were performed on the grid in combination with the def2-TZVP[59,88] basis set (def2-TZVP(-f) for LPNO-CEPA/1, 1296 basis functions, 4500 auxiliary basis functions). In both cases, we also performed calculations to correct for the basis set superposition error (BSSE). In Figure 8, the resulting potential energy surface for the BP86 case is shown.

Obviously, BP86[80] correctly predicts the location of the minimum for the hydrogen bond distance coordinate but completely fails for $\pi-\pi$ interactions. In Figure 9, the analogous graphs are shown for LPNO-CEPA/1. Pleasingly, clear minima almost exactly at the experimental geometry are found for both, the hydrogen bond and the $\pi-\pi$ interaction coordinate. In Figure 10, one-dimensional cuts are presented for both coordinates while keeping the other at the experimental value.

It appears that the hydrogen bond is calculated ∼0.1 Å shorter than the experimental value which appears to be within the experimental error and the resolution of the grid that was used to scan the energy in our calculations. For the case of $\pi-\pi$ interactions, the calculated minimum almost exactly corresponds with the experimental one.

We note in passing that the hydrogen bond energy was found to be 4.0 kcal/mol and the $\pi-\pi$ stacking energy −17.7 kcal/mol, both of which are plausible values. Obviously, our calculations neglect any impact of static disorder or crystal packing effects and they cannot be compared to experimental data. Nevertheless, this application should provide a reasonable feeling for what is presently routinely possible with the LPNO-CEPA/1 method.

## Conclusions

In this work, a parallel implementation of the LPNO-CEPA and LPNO-QCISD/LPNO-CCSD family of methods was presented. A detailed analysis of the parallelization scheme was presented together with representative timings. It was argued that the scaling of the algorithm is inherently nearly linear with the number of processors but that less than

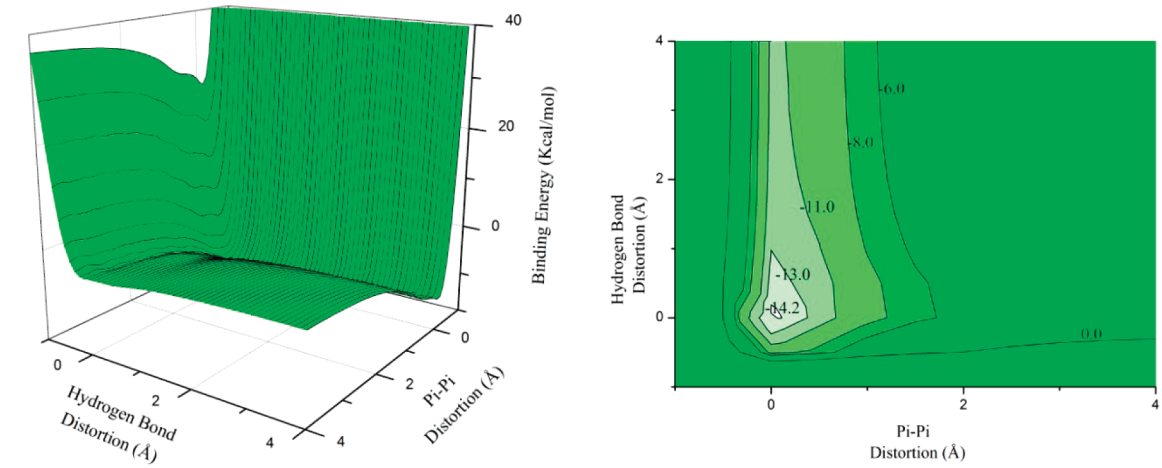



***Figure 9.*** The PES calculated for **2** with LPNO-CEPA/1/def2-TZVP(-f). Left: 3-dimensional plot, Right: Contour plot.

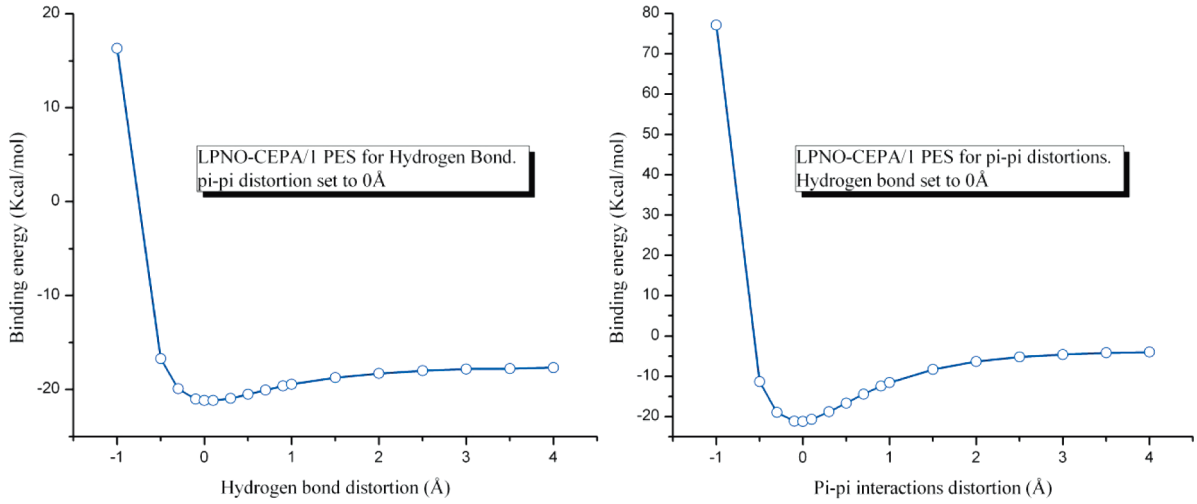



***Figure 10.*** One dimensional plots for hydrogen bond and $\pi-\pi$ interactions for the trimer calculated with LPNO-CEPA/1/def2-TZVP(-f). On the left we set $\pi-\pi$ equal to the experimental value and vary the Hydrogen bond distance while on the left we set the Hydrogen bond distance equal to the experimental one and distort the $\pi-\pi$ interaction distance.

optimum scaling is observed on present day machines due to limited memory bandwidth. Nevertheless, even with clusters built from cheap mass-market computers, speedups of about an order of magnitude can be obtained with 14 processors.

Second, the accuracy of the LPNO-CEPA/1 method was tested for weak intermolecular interactions. It was found that LPNO-CEPA/1 is an accurate method for such applications: the Mean Absolute Error was found to be 0.24 kcal/mol. This shows that no essential physics are lost with the LPNO approximations and that the canonical or LPNO-CEPA/1 methods provide results that are very close to CCSD(T) for such interactions. Given that LPNO-CEPA/1 is applicable to much larger systems than CCSD(T), this demonstrates a high potential of this method for chemical applications involving weak intermolecular interactions. Quite pleasingly, it was found that the accuracy of LPNO-CEPA/1 exceeds that of even the most purpose specific density functionals. The applicability of the LPNO-CEPA/1 method to larger systems was proven for the case of an acetophenone derivative trimer in which both hydrogen bonding and $\pi-\pi$ stacking interactions are prominent. This shows that large scale applications with ~1500 basis functions are rendered a routine application with the parallel version of the LPNO-CEPA/1 program. Together with the fact that the LPNO-CEPA/1 method is as easy to use as its canonical counterpart or a standard DFT functional, we do consider these developments as significant progress. A detailed discussion of the relative merits of LPNO-CEPA/1 and other local correlation schemes has been given in refs 48 and 49.

## References

(1) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, 2000.

(2) Neese, F. *Coord. Chem. Rev.* **2009**, *253*, 526–563.

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

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

(5) Peverati, R.; Baldridge, K. K. *J. Chem. Theory Comput.* **2008**, *4*, 2030–2048.

(6) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104–154119.

(7) Wodrich, M. D.; Jana, D. F.; Schleyer, P. v. R.; Corminboeuf, C. m. *J. Phys. Chem. A* **2008**, *112*, 11495–11500.

(8) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.

(9) Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *132*, 164113–164116.

(10) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104–224112.

(11) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2009**, *131*, 174105–174113.

(12) Román-Pérez, G.; Soler, J. M. *Phys. Rev. Lett.* **2009**, *103*, 096102.

(13) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *3*, 289–300.

(14) Kossmann, S.; Neese, F. *J. Chem. Theory Comput.* **2010**, *6*, 2325–2338.

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

(16) Gerenkamp, M.; Grimme, S. *Chem. Phys. Lett.* **2004**, *392*, 229–235.

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

(18) Lochan, R. C.; Head-Gordon, M. *J. Chem. Phys.* **2007**, *126*, 164101–164111.

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

(20) Lambrecht, D. S.; Doser, B.; Ochsenfeld, C. *J. Chem. Phys.* **2005**, *123*, 184102–184111.

(21) Doser, B.; Lambrecht, D. S.; Kussmann, J.; Ochsenfeld, C. *J. Chem. Phys.* **2009**, *130*, 064107−064114.

(22) Schutz, M.; Werner, H.-J.; Lindh, R.; Manby, F. R. *J. Chem. Phys.* **2004**, *121*, 737–750.

(23) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–262.

(24) Shavitt, I.; Bartlett, R. J. *Many-Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory*; Cambridge University Press: New York, 2009.

(25) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368–1377.

(26) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(27) Lee, T. J.; Rendell, A. P.; Taylor, P. R. *J. Phys. Chem.* **1990**, *94*, 5463–5468.

(28) Janowski, T.; Pulay, P. *J. Chem. Theory Comput.* **2008**, *4*, 1585–1592.

(29) Neese, F.; Hansen, A.; Wennmohs, F.; Grimme, S. *Acc. Chem. Res.* **2009**, *42*, 641–648.

(30) Meyer, W. *Int. J. Quantum Chem.* **1971**, *S5*, 341–348.

(31) Meyer, W. *J. Chem. Phys.* **1973**, *58*, 1017–1035.

(32) Meyer, W. *Theor. Chim. Acta* **1974**, *35*, 277–292.

(33) Ahlrichs, R.; Driessler, F.; Lischka, H.; Staemmler, V.; Kutzelnigg, W. *J. Chem. Phys.* **1975**, *62*, 1235–1247.

(34) Ahlrichs, R.; Driessler, F. *Theor. Chim. Acta* **1975**, *36*, 275–287.

(35) Schutz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691–5705.

(36) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286–6297.

(37) Mata, R. A.; Werner, H.-J. *J. Chem. Phys.* **2006**, *125*, 184110–184119.

(38) Maslen, P. E.; Head-Gordon, M. *Chem. Phys. Lett.* **1998**, *283*, 102–108.

(39) Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074116−074122.

(40) Maslen, P. E.; Lee, M. S.; Head-Gordon, M. *Chem. Phys. Lett.* **2000**, *319*, 205–212.

(41) Scuseria, G. E.; Ayala, P. Y. *J. Chem. Phys.* **1999**, *111*, 8330–8343.

(42) Ayala, P. Y.; Scuseria, G. E. *J. Comput. Chem.* **2000**, *21*, 1524–1531.

(43) Venkatnathan, A.; Szilva, A. B.; Walter, D.; Gdanitz, R. J.; Carter, E. A. *J. Chem. Phys.* **2004**, *120*, 1693–1704.

(44) Walter, D.; Venkatnathan, A.; Carter, E. A. *J. Chem. Phys.* **2003**, *118*, 8127–8139.

(45) Auer, A. A.; Nooijen, M. *J. Chem. Phys.* **2006**, *125*, 024104−024114.

(46) Edmiston, C.; Krauss, M. *J. Chem. Phys.* **1965**, *42*, 1119–1120.

(47) Meyer, W. *Configuration Expansion by Means of Pseudo-natural Orbitals. In Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum Press: New York, 1977; Vol. 3, pp 413−445.

(48) Neese, F.; Wennmohs, F.; Hansen, A. *J. Chem. Phys.* **2009**, *130*, 114108–114118.

(49) Neese, F.; Hansen, A.; Liakos, D. G. *J. Chem. Phys.* **2009**, *131*, 064103−064115.

(50) Pulay, P.; Saebo, S.; Meyer, W. *J. Chem. Phys.* **1984**, *81*, 1901–1905.

(51) Scuseria, G. E.; Janssen, C. L.; Schaefer III, H. F. *J. Chem. Phys.* **1988**, *89*, 7382–7387.

(52) Scuseria, G. E.; Schaefer III, H. F. *J. Chem. Phys.* **1989**, *90*, 3700–3703.

(53) Message Passing Interface Forum. www.mpi-forum.org (accessed Oct 15, 2010).

(54) Neese, F.; Becker, U.; Ganyushin, D.; Hansen, A.; Liakos, D. G.; Kollmar, C.; Kossmann, S.; Petrenko, T.; Reimann, C.; Riplinger, C.; Sivalingam, K.; Valeev, E.; Wezisla, B.; Wennmohs, F. *ORCA*; University of Bonn: Bonn, Germany, 2009.

(55) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

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

(57) Tewari, A. K.; Srivastava, P.; Puerta, C.; Valegra, P. *J. Mol. Struct.* **2009**, *921*, 251–254.

(58) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(59) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(60) Barker, J. K.; Davis, K.; Hoisie, A.; Kerbyson, J. D.; Lang, M.; Pakin, S.; Carlos, S. J. *In IEEE LSPP'08, LA-UR 07−6855*, 2008.

(61) Barker, J. K.; Davis, K.; Hoisie, A.; Kerbyson, J. D.; Lang, M.; Pakin, S.; Carlos, S. J. *Parallel Proc. Lett.* **2008**, *18*.

(62) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chem. Phys.* **2009**, *356*, 98–109.

(63) Muller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143–168.

(64) Sponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.

(65) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, 1997.

(66) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.

(67) Janowski, T.; Ford, A. R.; Pulay, P. *Mol. Phys.* **2010**, *108*, 249–257.

(68) Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.

(69) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *6*, 107–126.

(70) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *10*, 2747–2757.

(71) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 4061–4067.

(72) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *6*, 107–126.

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

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

(75) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585.

(76) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104–5.

(77) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

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

(79) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

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

(81) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(82) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

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

(84) Prof. Stefan Grimme Research Web Site. http://www.uni-muenster.de/Chemie.oc/grimme/en/index.html. (accessed July 6, 2010).

(85) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108−034116.

(86) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(87) Seth, S. K.; Hazra, D. K.; Mukherjee, M.; Kar, T. *J. Mol. Struct.* **2009**, *936*, 277–282.

(88) Schaefer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

CT100445S

JCTC Journal of Chemical Theory and Computation

# Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions

Kanchana S. Thanthiriwatte, Edward G. Hohenstein, Lori A. Burns, and C. David Sherrill*

*Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry and School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Received August 18, 2010

**Abstract:** Noncovalent interactions such as hydrogen bonds, van der Waals forces, and $\pi-\pi$ interactions play important roles influencing the structure, stability, and dynamic properties of biomolecules including DNA and RNA base pairs. In an effort to better understand the fundamental physics of hydrogen bonding (H-bonding), we investigate the distance dependence of interaction energies in the prototype bimolecular complexes of formic acid, formamide, and formamidine. Potential energy curves along the H-bonding dissociation coordinate are examined both by establishing reference CCSD(T) interaction energies extrapolated to the complete basis set limit and by assessing the performance of the density functional methods B3LYP, PBE, PBE0, B970, PB86, M05-2X, and M06-2X and empirical dispersion corrected methods B3LYP-D3, PBE-D3, PBE0-D3, B970-D2, BP86-D3, and $\omega$B97X-D, with basis sets 6-311++G(3df,3pd), aug-cc-pVDZ, and aug-cc-pVTZ. Although H-bonding interactions are dominated by electrostatics, it is necessary to properly account for dispersion interactions to obtain accurate energetics. In order to quantitatively probe the nature of hydrogen bonding interactions as a function of distance, we decompose the interaction energy curves into physically meaningful components with symmetry-adapted perturbation theory (SAPT). The SAPT results confirm that the contribution of dispersion and induction are significant at and near equilibrium, although electrostatics dominate. Among the DFT/DFT-D techniques, the best overall results are obtained utilizing counterpoise-corrected $\omega$B97X-D with the aug-cc-pVDZ basis set.

## 1. Introduction

Hydrogen-bonding (H-bonding) interactions, in addition to $\pi-\pi$ stacking interactions, provide a significant contribution to the stability and conformational arrangement of nucleic acids.[1−5] Though it is well understood that H-bonding arises mainly from electrostatic interactions,[6] the accurate determination of attendant interaction energies has been the subject of many theoretical studies.[1−23] While H-bonding between DNA/RNA base pairs is limited to O−H···N and N−H···N bonding patterns, the O−H···O motif is also accessible with nonstandard base pairs such as the uracil dimer.[22] As a step to better understanding hydrogen bonding in biomolecules, bimolecular complexes of formic acid (FaOO), formamide (FaON), and formamidine (FaNN) have been studied as prototypes for H-bonding in DNA/RNA base pairs.

Dissociation potential energy curves of the H-bonded complexes are governed not only by the dominant electrostatic component, but also by dispersion interactions near equilibrium and at medium distances. While it is well established that Hartree−Fock (HF) and density functional theory (DFT) generally recover most of the interaction energies in the H-bonded systems, these methods fail to

* To whom correspondence should be addressed: E-mail: sherrill@gatech.edu.

10.1021/ct100469b 
Published on Web 12/02/2010

describe dispersion interactions accurately.[24−29] Reliable computation of noncovalent interactions requires high-level treatment of electron correlation by methods such as coupled-cluster theory with singles and doubles including perturbative triples, CCSD(T).[30] This last approach has emerged as the most accurate computationally affordable method that can be applied to small systems and is considered the "gold standard" for chemical accuracy,[31] though it is unfortunately limited by its $O(N^7)$ complexity in its standard form. To describe noncovalent interactions in large systems, less computationally expensive methods must be employed.

Density functional theory[32] is used extensively to investigate a variety of chemical systems, as it generally requires fewer (for nonhybrid functionals) or equivalent (for hybrid functionals) computer resources compared to Hartree−Fock theory and provides reasonably accurate results.[32] The application of DFT to noncovalent interactions, however, has been hampered by the failure of most density functionals to describe "long-range" electron correlation.[24−29] Several approaches exist for improving existing density functionals. The addition of empirical terms to model dispersion interactions (i.e., the DFT-D method) is the most popular approach.[33−39] The hybrid meta exchange correlation functionals developed by Truhlar and co-workers are also promising candidates.[40−43] Truhlar's M05-2X and M06-2X functionals are said to account for "medium-range" electron correlation, which is sufficient for describing the dispersion interactions within many smaller complexes near their equilibrium geometries.[43]

We have examined potential energy curves of the FaOO−FaOO, FaON−FaON, FaNN−FaNN, FaOO−FaON, FaON−FaNN, and FaOO−FaNN bimolecular complexes, which we will refer to as the HBC6 test set. We present benchmark-quality complete-basis-set (CBS) extrapolated CCSD(T) potential energy curves for these molecules and also assess the performance of popular DFT and DFT-D methods. Previously, most investigations characterizing H-bonded complexes have focused on equilibrium configurations (from either geometry optimizations or crystal structures).[3,21] Recently, Hobza and co-workers also studied H-bonded complexes (formamide dimer, methylamine dimer, and methanol dimer) along the dissociation pathway.[5] The present work significantly extends these computations by examining five new H-bonded complexes, larger basis sets for the coupled-cluster correction, a much more extensive array of DFT methods, and wavefunction-based symmetry-adapted perturbation theory (SAPT) analysis to gain additional insight into the physical basis for attractive interactions in our target complexes. The systematic study performed here should be very helpful in elucidating the fundamental nature of H-bonding interactions (including their distance dependence) and in providing additional benchmark CCSD(T)/CBS data for H-bonded systems, which we have found necessary in independent work on parametrization of dispersion-including DFT methods.

## 2. Theoretical Methods

Geometry optimizations were performed for the homogeneous and heterogeneous bimolecular complexes of FaOO, FaON, and FaNN in planar hydrogen-bonded configurations. The structures of these dimers along the dissociation coordinate were fully optimized at the CCSD(T)/aug-cc-pVDZ[44,45] level of theory under the constraint of fixed intermonomer distances. We defined the dissociation coordinate for our systems as the distance between the central carbon atoms of each monomer in the bimolecular complexes. It is noteworthy that most of the complexes in HBC6, particularly FaOO−FaOO and FaOO−FaNN, show double proton transfer with "short, strong" or "low-barrier" hydrogen bonds.[46−52] Because our interest here is primarily in noncovalent interactions and not in proton transfer reactions, we have neglected very short intermolecular distances where the proton transfer occurs. Single-point energy computations along the dissociation coordinate were performed with second order Møller−Plesset perturbation theory (MP2) using Dunning's aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ basis sets,[44,45] as well as CCSD(T) with the aug-cc-pVDZ and aug-cc-pVTZ basis sets.

Estimates of the CBS limit of the CCSD(T) interaction energies were obtained first by extrapolating to the CBS MP2 limit using the two-point extrapolation scheme of Halkier et al.[53] with aug-cc-pVTZ and aug-cc-pVQZ basis sets. Next, we added a "coupled-cluster correction", ΔCCSD(T), estimated as the difference between CCSD(T) and MP2 in a smaller basis set. Janowski and Pulay[54] have emphasized that basis sets beyond aug-cc-pVDZ can be important for high-quality estimates of ΔCCSD(T), and we have recently confirmed this for several other van der Waals dimers.[16,21] In this work, the estimates of ΔCCSD(T) were evaluated using aug-cc-pVDZ/aug-cc-pVTZ extrapolations of the MP2 and CCSD(T) correlation energies. The final CCSD(T)/CBS estimation was obtained by summing the energies of Hartree−Fock/aug-cc-pVQZ, the extrapolated MP2/CBS correlation, and the estimated ΔCCSD(T) correction. All CCSD(T) and MP2 computations were performed with the core electrons frozen.

All geometry optimizations were performed with the ACES II program suite,[55] and single-point energy computations utilized the Molpro2009 package of ab initio programs.[56] DFT computations along each curve were performed using the Q-Chem 3.2 suite of programs[57] utilizing a Lebedev grid with 302 angular points for each of the 100 radial shells. This grid is larger than the default of most electronic structure program packages, but it is necessary to avoid artifacts due to numerical integration for noncovalent interactions, particularly when using meta-GGA functionals.[24,58] Energy evaluations were performed for the B3LYP,[59] PBE,[60] PBE0,[61] B970,[62] BP86,[63,64] M05-2X,[40] M06-2X,[41,42] and ωB97X-D[39,65] functionals along with the 6-311++G(3df,3pd), aug-cc-pVDZ, and aug-cc-pVTZ basis sets. Here, we denote the original B97 functional as B970 to distinguish it from the reparameterized B97 functional developed as part of the B97-D method of Grimme,[37] which has previously been shown to perform well for potential curves of more weakly bound van der Waals dimers.[17] The empirical dispersion correction (denoted -D2 and -D3) introduced by Grimme[37,38] was added to B3LYP, PBE, PBE0, B970, and BP86 functionals. The dispersion interaction energy ($E_{disp}$) is given by

$$E_{\mathrm{disp}} = -\sum_{n}\sum_{\mathrm{AB}} s_n \frac{C_n^{\mathrm{AB}}}{R_{\mathrm{AB}}^n} f_{\mathrm{d},n}(R_{\mathrm{AB}}), n = 6, 8 \quad (1)$$

where $s_n$ is a global scaling factor, $C_n^{\mathrm{AB}}$ denotes the averaged $n$th-order dispersion coefficient for atom pair AB, $R_{\mathrm{AB}}$ is interatomic distance of atom pair AB, and $f_{\mathrm{d},n}(R_{\mathrm{AB}})$ is the damping function. The use of a damping function minimizes double counting of "short-range" correlation effects captured by the density functional. In the -D2 approach, the damping function is given by

$$f_{\mathrm{d},n}(R_{\mathrm{AB}}) = \frac{1}{1 + \mathrm{e}^{-d(R_{\mathrm{AB}}/R_{\mathrm{r}}-1)}}, n = 6 \quad (2)$$

where $R_{\mathrm{r}}$ represents the sum of atomic radii and $d$ is a global damping parameter that controls the "sharpness" of the damping function. The $R^{-6}$ term is the leading term in the expansion of the dispersion energy of which higher order ($R^{-8}$, $R^{-10}$, ...) terms are omitted. The empirical dispersion energies were scaled by a global factor, $s_6$, which is 1.05 for B3LYP, 0.75 for PBE, 0.6 for PBE0, and 1.05 for BP86, as recommended by Grimme[37] and 0.75 for B970.[66,67] The damping function of the -D3 approach is given by

$$f_{\mathrm{d},n}(R_{\mathrm{AB}}) = \frac{1}{1 + 6(R_{\mathrm{AB}}/s_{\mathrm{r},n}R_{\mathrm{AB}}^0)}, n = 6, 8 \quad (3)$$

where $s_{\mathrm{r},n}$ is the order-dependent scaling factor of the cutoff radii $R_{\mathrm{AB}}^0$. The -D3 correction was added to the B3LYP, PBE, PBE0, and BP86 functionals. Chai and Head-Gordon's[39] $\omega$B97X-D functional incorporates a similar $R^{-6}$ dispersion term to improve its performance for treating noncovalent complexes. The dispersion corrected functionals will be referred to as either DFT-D2 or DFT-D3 appropriately. Grimme's DFT-D2 empirical dispersion correction was implemented in Q-Chem 3.2 by two of the authors (C.D.S. and E.G.H.).

To reduce basis set superposition error, we applied the Boys−Bernardi scheme[68] to yield counterpoise-corrected interaction energies, $E_{\mathrm{cp}}^{\mathrm{int}}$. We also obtained "relaxed" interaction energies, $E_{\mathrm{rlx}}^{\mathrm{int}}$, by adding to $E_{\mathrm{cp}}^{\mathrm{int}}$ the deformation energies:

$$E_{\mathrm{rlx}}^{\mathrm{int}} = E_{\mathrm{cp}}^{\mathrm{int}} + (E_{\mathrm{A}}^{\mathrm{A}} - E_{\mathrm{A}'}^{\mathrm{A}}) + (E_{\mathrm{B}}^{\mathrm{B}} - E_{\mathrm{B}'}^{\mathrm{B}}) \quad (4)$$

where $E_{\mathrm{A}}^{\mathrm{A}} - E_{\mathrm{A}'}^{\mathrm{A}}$ represents the deformation energy required to take monomer A from its equilibrium geometry to that in the complex (and, as denoted by the superscript, this energy difference is evaluated in the basis of monomer A). Thus, $E_{\mathrm{rlx}}^{\mathrm{int}}$ denotes the counterpoise-corrected interaction energy of the dimer relative to infinitely separated monomers in their equilibrium geometries, whereas $E_{\mathrm{cp}}^{\mathrm{int}}$ denotes the counterpoise-corrected interaction energy of the dimer relative to infinitely separated monomers frozen at the monomer geometries they adopt in the dimer.

The interaction energies of the six H-bonded dimers were computed with symmetry-adapted perturbation theory.[69,70] SAPT computes the interaction in terms of individual energy components. For the purpose of our analysis, we will group the individual terms as

$$E_{\mathrm{electrostatic}} = E_{\mathrm{elst}}^{(10)} + E_{\mathrm{elst,resp}}^{(12)} + E_{\mathrm{elst,resp}}^{(13)} \quad (5)$$

$$E_{\mathrm{exchange}} = E_{\mathrm{exch}}^{(10)} + E_{\mathrm{exch}}^{(11)} + E_{\mathrm{exch}}^{(12)} \quad (6)$$

$$E_{\mathrm{induction}} = E_{\mathrm{ind,resp}}^{(20)} + E_{\mathrm{exch-ind,resp}}^{(20)} + {}^{t}E_{\mathrm{ind}}^{(22)} + {}^{t}E_{\mathrm{exch-ind}}^{(22)} + \delta E_{\mathrm{ind,resp}}^{(\mathrm{HF})} \quad (7)$$

$$E_{\mathrm{dispersion}} = E_{\mathrm{disp}}^{(20)} + E_{\mathrm{disp}}^{(30)} + E_{\mathrm{disp}}^{(21)} + E_{\mathrm{disp}}^{(22)} + E_{\mathrm{exch-disp}}^{(20)} \quad (8)$$

The SAPT computations have been performed with our recently developed SAPT program.[71] This program has been implemented within the framework of PSI 3.4.[72] The level of SAPT shown above, which we will denote SAPT2+(3), when paired with the aug-cc-pVDZ basis, can be expected to provide energies within 0.4 kcal mol$^{-1}$ of benchmark results for systems of this size.[73] All SAPT computations have been performed under the density fitting (DF) approximation using the auxiliary aug-cc-pVDZ-RI basis.[74]

## 3. Results and Discussion

The performance of density functional methods for H-bonded interaction energies at different distances is illustrated in Figure 1. Relaxed, CP-corrected interaction energies vs dissociation coordinate distance have been plotted for each of the six complexes of HBC6 using an aug-cc-pVTZ basis set in combination with B3LYP-D3, PBE-D3, PBE0-D3, BP86-D3, B970-D2, M05-2X, M06-2X, and $\omega$B97X-D and their underlying uncorrected DFT functionals where applicable. The interaction energy curves for B3LYP-D2, PBE-D2, PBE0-D2, and BP86-D2 can be found in the Supporting Information. The CCSD(T)/CBS curves are also presented for reference. Although standard DFT methods are often thought to be reliable for describing H-bonding interactions, it can be clearly observed that DFT methods B970, B3LYP, and BP86 are significantly underbound, with average absolute errors (MAEs) at the potential minima of 1.2−2.0 kcal mol$^{-1}$. In contrast, PBE-D3, BP86-D3, PBE0-D3, and B3LYP-D3 approaches are overbound with respect to the reference energies with MAE values of 1.6−1.8 kcal mol$^{-1}$. Surprisingly, the uncorrected PBE and PBE0 functionals exhibit far better performance than their DFT-D counterparts, with MAEs of 0.43 and 0.35 kcal mol$^{-1}$ at their equilibrium geometries. These two are the most accurate DFT treatments, along with B970-D2, M05-2X, $\omega$B97X-D, and M06-2X, all of which yield potential minimum MAE values less than 0.5 kcal mol$^{-1}$. All methods display the correct asymptotic behavior at the dissociation limit.

Mean absolute error statistics for $E_{\mathrm{rlx}}^{\mathrm{int}}$ with respect to CCSD(T)/CBS values across the entire potential energy curves are compiled in Table 1 for the functionals examined. All three basis sets show similar performance. The DFT methods B970, B3LYP, and BP86 as well as the dispersion corrected DFT methods PBE-D3, PBE0-D3, BP86-D3, and their -D2 counterparts consistently give overall MAE values (weighted average over six test cases, rightmost columns) greater than 0.9 kcal mol$^{-1}$ for all basis sets. M05-2X, B970-D, $\omega$B97X-D, and M06-2X have errors reliably less than 0.4 kcal mol$^{-1}$. The best-performing functional (for relaxed interaction energies) is M06-2X, with $\omega$B97X-D the second-best. The -D3 MAE for both the PBE0 and BP86 functionals

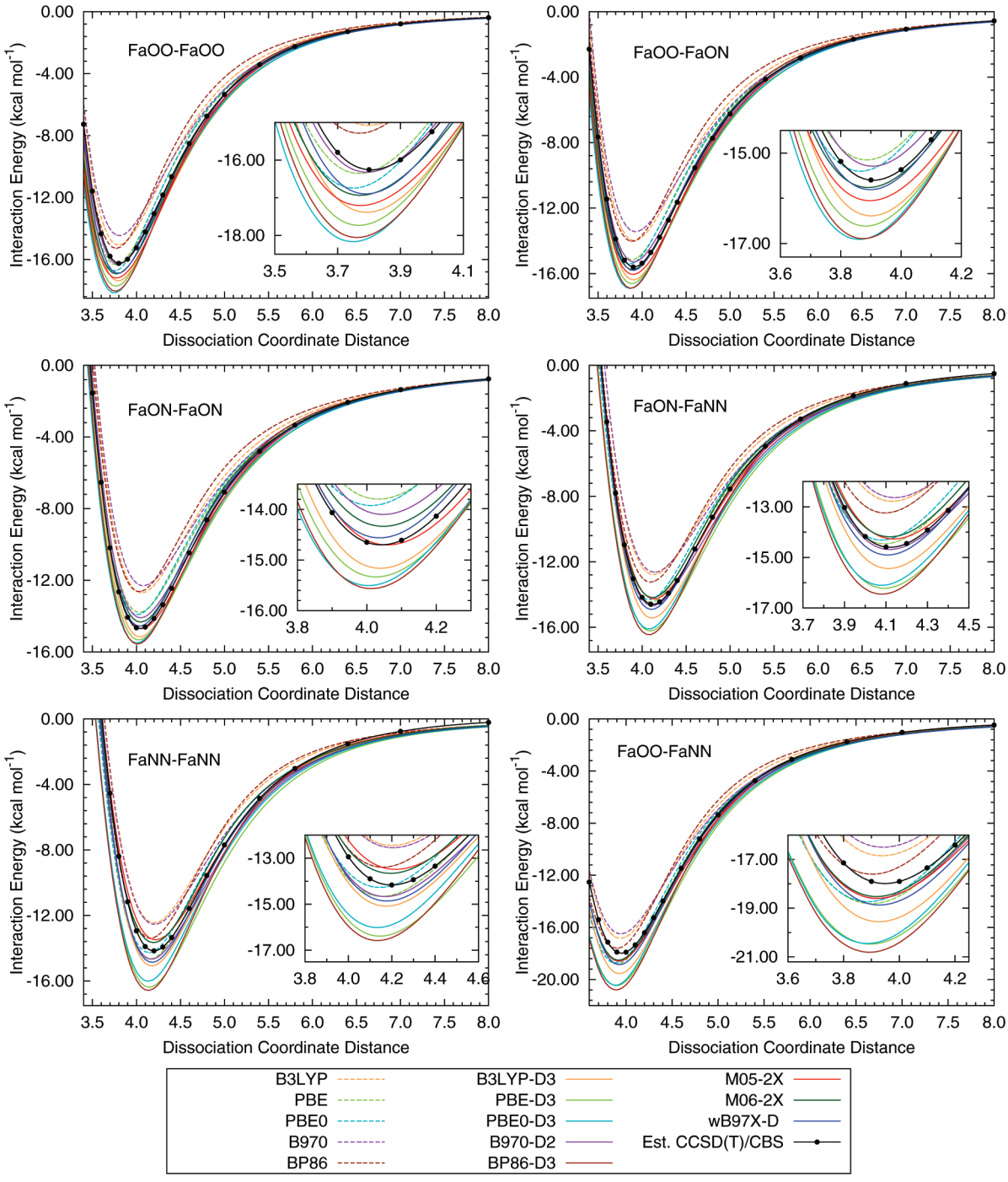



***Figure 1.*** Relaxed counterpoise-corrected interaction energies at the DFT(-D)/aug-cc-pVTZ and CCSD(T)/CBS levels of theory.

is about 0.06−0.1 kcal mol$^{-1}$ higher than the -D2 error. PBE-D3 and B3LYP-D3 show a MAE value of 0.1−0.2 kcal mol$^{-1}$ improvement in reliability compared with their -D2 counterparts, although PBE-D3 has an error higher than 1 kcal mol$^{-1}$. We note that the -D3 correction to the popular B3LYP functional leads to overall MAEs of around 0.6 kcal mol$^{-1}$, which is not as good as the meta-GGA functionals, $\omega$B97X-D, or B970-D2, but better than the others.

It is also worthwhile to examine the counterpoise-corrected interaction energies, $E_{cp}^{int}$ (without the addition of deformation energies), because these are somewhat more convenient for use as benchmarks for assessing new theoretical methods. The MAE statistics are tabulated in Table 2, and errors with respect to the CCSD(T)/CBS reference are plotted in Figure 2 (negative values indicate overbinding). Our results show a considerable increase in the error at intermolecular distances shorter than the equilibrium distance. Most methods are overbound at closer intermonomer distances except B3LYP, B970, and BP86, which are highly underbound. However, the dispersion corrected B970-D functional is the second best performing functional with the 6-311++G(3df,3pd) and aug-cc-pVTZ basis sets, and the third best with the aug-cc-pVDZ basis set. The DFT/DFT-D methods B3LYP,

***Table 1.*** Errors in DFT(-D) Relaxed Interaction Energies with Respect to CCSD(T)/CBS[a]

| | mean absolute error (MAE) in kcal mol$^{-1}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | FaOO−FaOO | FaON−FaON | FaNN−FaNN | FaOO−FaON | FaON−FaNN | FaOO−FaNN | HBC6 | HBC6(D2) |
| | | | | 6-311++G(3df,3pd) | | | | |
| B970 | 0.95 | 1.59 | 1.02 | 1.27 | 1.31 | 0.83 | 1.17 | |
| B3LYP | 0.77 | 1.41 | 1.22 | 1.07 | 1.31 | 0.76 | 1.10 | |
| BP86 | 0.86 | 1.44 | 0.71 | 1.13 | 0.87 | 0.68 | 0.95 | |
| PBE0 | 0.50 | 0.44 | 0.78 | 0.44 | 0.51 | 0.50 | 0.53 | |
| PBE | 0.36 | 0.55 | 0.80 | 0.32 | 0.34 | 0.44 | 0.47 | |
| PBE0-D3 | 1.07 | 0.69 | 1.75 | 0.88 | 1.23 | 1.31 | 1.15 | 1.09 |
| PBE-D3 | 0.90 | 0.51 | 2.01 | 0.71 | 1.27 | 1.42 | 1.13 | 1.35 |
| BP86-D3 | 0.91 | 0.53 | 1.96 | 0.71 | 1.24 | 0.26 | 1.12 | 1.03 |
| B3LYP-D3 | 0.68 | 0.27 | 0.71 | 0.48 | 0.50 | 0.87 | 0.58 | 0.71 |
| M05-2X | 0.54 | 0.20 | 0.42 | 0.26 | 0.29 | 2.33 | 0.33 | |
| B970-D2 | 0.16 | 0.43 | 0.67 | 0.24 | 0.19 | 0.40 | 0.35 | |
| $\omega$B97X-D | 0.33 | 0.15 | 0.60 | 0.14 | 0.26 | 0.51 | 0.33 | |
| M06-2X | 0.36 | 0.33 | 0.41 | 0.24 | 0.35 | 0.27 | 0.33 | |
| | | | | aug-cc-pVDZ | | | | |
| B970 | 1.00 | 1.53 | 1.01 | 1.27 | 1.28 | 0.81 | 1.16 | |
| B3LYP | 0.80 | 1.34 | 1.17 | 1.06 | 1.25 | 0.72 | 1.06 | |
| BP86 | 0.84 | 1.39 | 0.62 | 1.12 | 0.82 | 0.62 | 0.91 | |
| PBE0 | 0.40 | 0.30 | 0.69 | 0.31 | 0.39 | 0.45 | 0.42 | |
| PBE | 0.27 | 0.47 | 0.84 | 0.24 | 0.26 | 0.46 | 0.42 | |
| PBE0-D3 | 1.05 | 0.79 | 1.78 | 0.92 | 1.30 | 1.36 | 1.20 | 1.14 |
| PBE-D3 | 0.90 | 0.60 | 2.06 | 0.74 | 1.34 | 1.49 | 1.18 | 1.40 |
| BP86-D3 | 0.89 | 0.57 | 1.98 | 0.72 | 1.35 | 1.41 | 1.15 | 1.03 |
| B3LYP-D3 | 0.65 | 0.39 | 0.76 | 0.50 | 0.56 | 0.91 | 0.62 | 0.75 |
| M05-2X | 0.38 | 0.38 | 0.50 | 0.14 | 0.41 | 0.24 | 0.34 | |
| B970-D2 | 0.13 | 0.44 | 0.69 | 0.28 | 0.19 | 0.44 | 0.36 | |
| $\omega$B97X-D | 0.34 | 0.16 | 0.66 | 0.18 | 0.35 | 0.56 | 0.37 | |
| M06-2X | 0.26 | 0.29 | 0.32 | 0.10 | 0.22 | 0.25 | 0.24 | |
| | | | | aug-cc-pVTZ | | | | |
| B970 | 0.93 | 1.55 | 0.97 | 1.23 | 1.26 | 0.79 | 1.13 | |
| B3LYP | 0.78 | 1.38 | 1.19 | 1.06 | 1.28 | 0.74 | 1.08 | |
| BP86 | 0.86 | 1.41 | 0.70 | 1.10 | 0.83 | 0.68 | 0.93 | |
| PBE0 | 0.54 | 0.43 | 0.80 | 0.46 | 0.52 | 0.53 | 0.54 | |
| PBE | 0.39 | 0.51 | 0.84 | 0.33 | 0.35 | 0.47 | 0.48 | |
| PBE0-D3 | 1.08 | 0.73 | 1.80 | 0.91 | 1.27 | 1.34 | 1.19 | 1.13 |
| PBE-D3 | 0.91 | 0.55 | 2.05 | 0.73 | 1.31 | 1.45 | 1.16 | 1.38 |
| BP86-D3 | 0.92 | 0.57 | 1.99 | 0.75 | 1.27 | 1.41 | 1.15 | 1.07 |
| B3LYP-D3 | 0.67 | 0.29 | 0.74 | 0.50 | 0.53 | 0.89 | 0.60 | 0.73 |
| M05-2X | 0.73 | 0.13 | 0.30 | 0.47 | 0.17 | 0.45 | 0.37 | |
| B970-D2 | 0.16 | 0.38 | 0.73 | 0.18 | 0.21 | 0.44 | 0.35 | |
| $\omega$B97X-D | 0.32 | 0.16 | 0.61 | 0.15 | 0.25 | 0.50 | 0.33 | |
| M06-2X | 0.48 | 0.20 | 0.37 | 0.25 | 0.27 | 0.34 | 0.32 | |

[a] MAEs are calculated across each potential energy curve. The column HBC6 denotes the MAEs over all six potential energy curves.

B970, BP86, PBE-D2/-D3, and PBE0-D2/-D3 show poor performance for H-bonded $E_{cp}^{int}$, similar to the results for $E_{rlx}^{int}$. M05-2X, M06-2X, B970-D2, and $\omega$B97X-D demonstrate MAE values of less than 0.4 kcal mol$^{-1}$, with $\omega$B97X-D and B970-D as the best performing. The MAE of -D3 in the unrelaxed interaction energies for BP86 and PBE0 is up to 0.08 kcal mol$^{-1}$ higher than -D2 with respect to CCSD(T)/CBS values. B3LYP-D3 and PBE-D3 have improved the reliability by 0.1−0.2 kcal mol$^{-1}$ from B3LYP-D2 and PBE-D2 functionals.

Our computations show that the DFT methods B3LYP, BP86, and B970 cannot describe the H-bonding interaction quantitatively. All three offer MAE values greater than 1.25 kcal mol$^{-1}$ and across the entire test set are significantly underbound with respect to the CCSD(T)/CBS reference. Although these methods should be able to describe the dominant electrostatic contributions, in order to obtain accurate interaction energies, the long- and medium-range dispersion contributions also need to be added. After careful observation of the B3LYP, B970, and BP86 error curves in Figure 2, one might suggest that they behave as $1/R^6$. B3LYP-D3 and BP86-D3 compensate for this deficiency (although they often overcompensate at short intermolecular distances), and B970-D2 shows excellent performance with MAE values less than 0.30 kcal mol$^{-1}$. Hence, in order to accurately represent interaction energies in H-bonded systems, the dispersion energy contribution needs to be included. Unlike B3LYP, BP86, and B970, our results show that standard PBE and PBE0 are fairly reliable across the potential energy curves for all three basis sets. The MAEs in interaction energies for PBE and PBE0 are 0.52 and 0.40 kcal mol$^{-1}$, respectively, with the aug-cc-pVTZ basis. Since pure PBE and PBE0 already behave well for the interaction energies of these systems, the addition of an empirical dispersion energy term makes PBE-D2/-D3 and PBE0-D2/-D3 worse for interaction energies in H-bonded systems. Interaction energy error curves for M06-2X are better than for M05-

***Table 2.*** Errors in DFT(-D) Unrelaxed Interaction Energies with Respect to CCSD(T)/CBS[a]

| | mean absolute error (MAE) in kcal mol$^{-1}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| method | FaOO−FaOO | FaON−FaON | FaNN−FaNN | FaOO−FaON | FaON−FaNN | FaOO−FaNN | HBC6 | HBC6(D2) |
| | | | | 6-311++G(3df,3pd) | | | | |
| BP86 | 1.24 | 1.80 | 1.02 | 1.55 | 1.40 | 1.13 | 1.36 | |
| B970 | 1.06 | 1.59 | 1.36 | 1.32 | 1.47 | 1.02 | 1.31 | |
| B3LYP | 0.90 | 1.44 | 1.56 | 1.16 | 1.49 | 1.01 | 1.26 | |
| PBE | 0.44 | 0.93 | 0.30 | 0.71 | 0.46 | 0.25 | 0.52 | |
| PBE0 | 0.40 | 0.43 | 0.45 | 0.34 | 0.33 | 0.37 | 0.39 | |
| PBE0-D3 | 0.97 | 0.67 | 1.31 | 0.82 | 1.01 | 1.09 | 0.98 | 0.90 |
| BP86-D3 | 0.51 | 0.25 | 1.27 | 0.35 | 0.76 | 0.79 | 0.65 | 0.61 |
| PBE-D3 | 0.47 | 0.17 | 1.24 | 0.28 | 0.70 | 0.78 | 0.60 | 0.80 |
| B3LYP-D3 | 0.56 | 0.28 | 0.35 | 0.41 | 0.32 | 0.60 | 0.42 | 0.61 |
| M05-2X | 0.63 | 0.14 | 0.55 | 0.36 | 0.29 | 0.22 | 0.37 | |
| M06-2X | 0.35 | 0.20 | 0.34 | 0.21 | 0.27 | 0.20 | 0.26 | |
| B970-D2 | 0.14 | 0.46 | 0.23 | 0.30 | 0.18 | 0.17 | 0.25 | |
| $\omega$B97X-D | 0.29 | 0.08 | 0.32 | 0.15 | 0.17 | 0.39 | 0.23 | |
| | | | | aug-cc-pVDZ | | | | |
| BP86 | 1.41 | 1.91 | 1.06 | 1.69 | 1.48 | 1.23 | 1.47 | |
| B970 | 1.23 | 1.69 | 1.41 | 1.47 | 1.55 | 1.11 | 1.42 | |
| B3LYP | 1.07 | 1.54 | 1.60 | 1.31 | 1.56 | 1.11 | 1.37 | |
| PBE | 0.59 | 1.01 | 0.29 | 0.83 | 0.51 | 0.32 | 0.59 | |
| PBE0 | 0.33 | 0.51 | 0.42 | 0.33 | 0.30 | 0.34 | 0.37 | |
| PBE0-D3 | 0.82 | 0.59 | 1.28 | 0.70 | 0.95 | 1.02 | 0.89 | 0.81 |
| BP86-D3 | 0.36 | 0.18 | 1.24 | 0.23 | 0.70 | 0.70 | 0.57 | 0.57 |
| PBE-D3 | 0.32 | 0.15 | 1.23 | 0.19 | 0.65 | 0.72 | 0.54 | 0.57 |
| B3LYP-D3 | 0.38 | 0.25 | 0.32 | 0.29 | 0.27 | 0.50 | 0.33 | 0.41 |
| M05-2X | 0.37 | 0.34 | 0.69 | 0.14 | 0.48 | 0.12 | 0.36 | |
| M06-2X | 0.19 | 0.31 | 0.35 | 0.16 | 0.32 | 0.16 | 0.25 | |
| B970-D2 | 0.28 | 0.54 | 0.19 | 0.42 | 0.25 | 0.10 | 0.30 | |
| $\omega$B97X-D | 0.18 | 0.12 | 0.29 | 0.10 | 0.14 | 0.32 | 0.19 | |
| | | | | aug-cc-pVTZ | | | | |
| BP86 | 1.25 | 1.81 | 0.98 | 1.56 | 1.39 | 1.12 | 1.35 | |
| B970 | 1.06 | 1.59 | 1.30 | 1.32 | 1.44 | 1.01 | 1.29 | |
| B3LYP | 0.92 | 1.45 | 1.53 | 1.18 | 1.48 | 1.02 | 1.27 | |
| PBE | 0.45 | 0.93 | 0.31 | 0.72 | 0.44 | 0.24 | 0.52 | |
| PBE0 | 0.43 | 0.44 | 0.46 | 0.36 | 0.34 | 0.39 | 0.40 | |
| PBE0-D3 | 0.96 | 0.66 | 1.36 | 0.81 | 1.03 | 1.10 | 0.98 | 0.91 |
| BP86-D3 | 0.51 | 0.27 | 1.32 | 0.36 | 0.78 | 0.81 | 0.67 | 0.64 |
| PBE-D3 | 0.46 | 0.16 | 1.29 | 0.27 | 0.72 | 0.79 | 0.61 | 0.81 |
| B3LYP-D3 | 0.53 | 0.27 | 0.39 | 0.39 | 0.33 | 0.59 | 0.41 | 0.51 |
| M05-2X | 0.75 | 0.13 | 0.44 | 0.48 | 0.20 | 0.34 | 0.39 | |
| M06-2X | 0.43 | 0.13 | 0.30 | 0.26 | 0.20 | 0.21 | 0.26 | |
| B970-D2 | 0.11 | 0.42 | 0.29 | 0.27 | 0.14 | 0.18 | 0.24 | |
| $\omega$B97X-D | 0.27 | 0.09 | 0.33 | 0.13 | 0.15 | 0.36 | 0.22 | |

[a] MAEs are calculated across each potential energy curve. The column HBC6 denotes the MAEs over all six potential energy curves.

2X, but both MAE values are under 0.40 kcal mol$^{-1}$ for all three basis sets. The best performing DFT/DFT-D functional for unrelaxed interaction energies of the HBC6 set is $\omega$B97X-D, which gives MAE values of 0.23, 0.19, and 0.22 kcal mol$^{-1}$ with 6-311++G(3df,3pd), aug-cc-pVDZ, and aug-cc-pVTZ, respectively.

The HBC6 test set shows that appropriately chosen DFT-D methods and meta-GGA methods can accurately describe noncovalent interactions in H-bonded systems. To further understand the nature of these H-bonding interactions, we decompose the interaction energy to its components (electrostatics, exchange, induction, and dispersion) as described in eqs 5−8. We use a ternary diagram to interpret the results of our SAPT computations.[75] In a ternary diagram, each of three quantities is represented by a vertex which denotes 100% of the considered quantity, and the opposite side of the triangle is 0% in that quantity. Any other value between 0% and 100% can be placed between the vertex for that quantity and the opposite side. Figure 3 shows the ternary diagram for the attractive components of the total interaction energies: electrostatics, induction, and dispersion. For each system in the HBC6 test set, we plot a dot for each intermolecular distance $R$. The dots that are closer to the electrostatics vertex belong to the large intermolecular distances, and dots that are closer to the middle of the diagram correspond to the shorter intermolecular distances for each system. The colored lines correspond to the SAPT attractive components at the minima of the FaOO−FaOO, FaON−FaON, and FaNN−FaNN systems. It is clear that electrostatics dominate at long range, taking over from induction and dispersion contributions, which account for 10−40% and 10−20%, respectively, of attraction at short to intermediate distances. This decomposition analysis agrees with our results, in which DFT methods generally fail to describe the interaction energies in H-bonded systems accurately, while DFT-D methods and meta-GGA

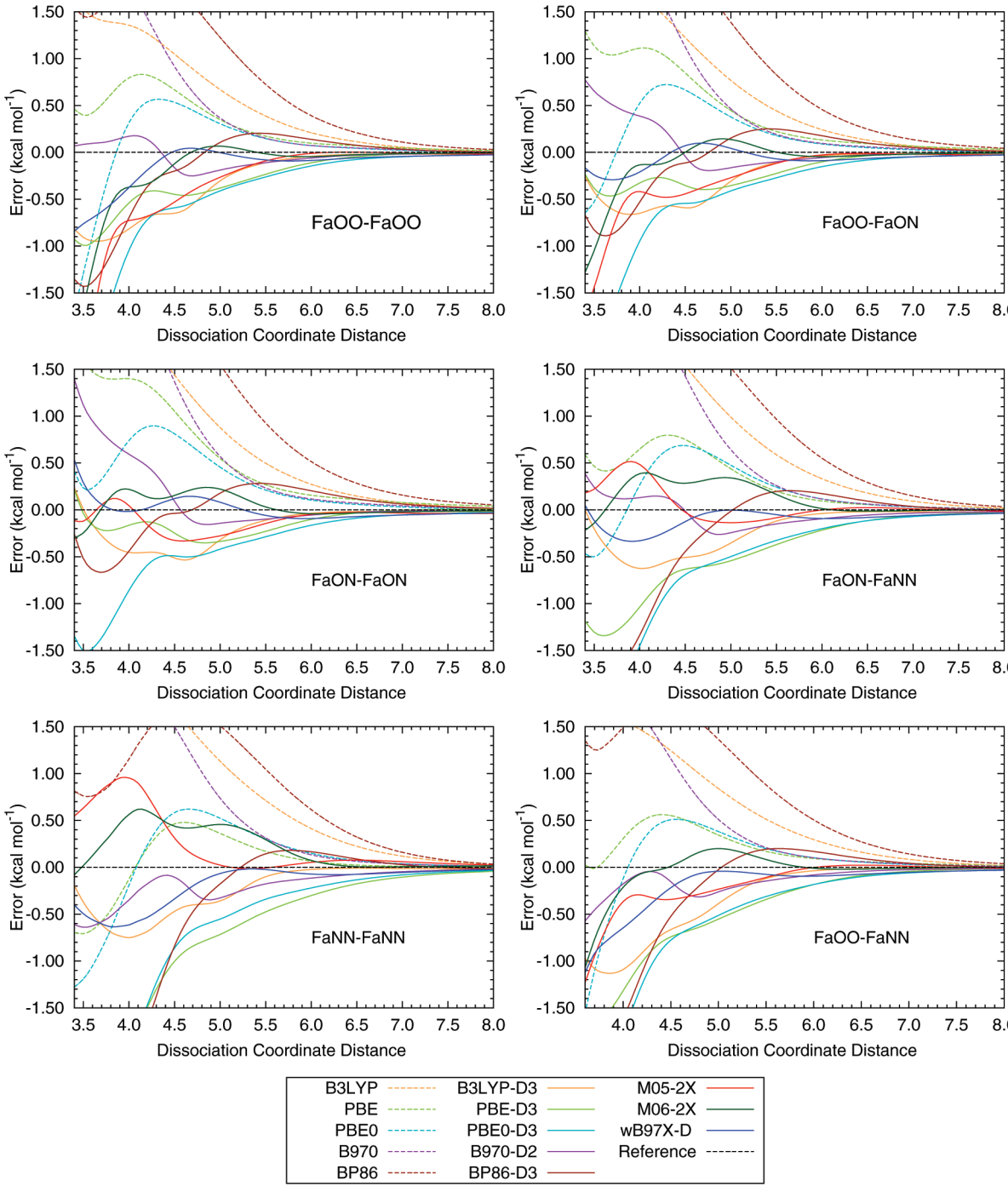



***Figure 2.*** Errors in counterpoise-corrected DFT(-D)/aug-cc-pVTZ unrelaxed interaction energies with respect to CCSD(T)/CBS.

methods that claim to recover "medium-range" electron correlation show success.

## 4. Conclusions

We have studied counterpoise corrected interaction energies (with and without deformation energy corrections) for the HBC6 test set of H-bonded complexes. The benchmark CCSD(T)/CBS potential curves reported here should be helpful in testing and parametrizing new approximate methods for noncovalent interactions and supplement existing benchmark data which is primarily available for potential energy minima only. DFT-D and meta-GGA DFT methods provide significant improvements over traditional density functionals for these systems. H-bonded complexes are dominated by electrostatic influences, and thus traditional density functionals are generally capable of computing reasonable binding energies. However, our results have shown that the DFT methods B3LYP, B970, and BP86 perform poorly for our test set, while PBE and PBE0 provide accurate results. The meta-GGA functionals and several of the DFT-D methods perform even better, and this improvement is

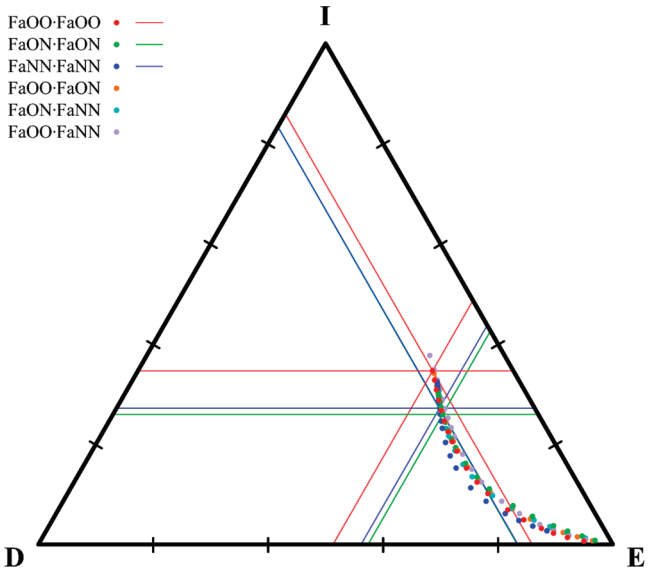



***Figure 3.*** Relationship between Electrostatics (E)−Dispersion (D)−Induction (I) in SAPT analysis for HBC6 test set. Proximity to a vertex indicates an increasing fraction of the attraction coming from that component. The dots that are closer to the electrostatics vertex belong to the large intermolecular distances, and dots that are closer to the middle of the diagram correspond to the shorter intermolecular distances for each system (see the text).

ascribed to the importance of dispersion interactions at medium range, as revealed by SAPT analysis. Thus, although H-bonded interactions are dominated by electrostatic interactions, the contribution of London dispersion forces is not negligible in computing accurate interaction energies.

**Acknowledgment.** This material is based upon work supported by the National Science Foundation through grant CHE-1011360.

**Supporting Information Available:** Errors in DFT(-D2/-D3) relaxed interaction energies, errors in DFT(-D2/-D3) unrelaxed interaction energies, Cartesian coordinates of optimized geometries, and counterpoise-corrected interaction energies for all considered DFT/DFT-D methods of the HBC6 test set. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Pitoňák, M.; Riley, K. E.; Neogrady, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636–1644.

(2) Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

(3) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(4) Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697.

(5) Riley, K. E.; Pitonak, M.; Cerny, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 66–80.

(6) Scheiner, S. *Hydrogen Bonding: A Theoretical Perspective*; Oxford Univ Press: New York, 1997.

(7) Frisch, M. J.; Del Bene, J. E.; Binkley, J. S.; Schaefer, H. F. *J. Chem. Phys.* **1986**, *84*, 2279–2289.

(8) Feller, D.; Boyle, C. M.; Davidson, E. R. *J. Chem. Phys.* **1987**, *86*, 3424–3440.

(9) Szalewicz, K.; Cole, S. J.; Kolos, W.; Bartlett, R. J. *J. Chem. Phys.* **1988**, *89*, 3662–3673.

(10) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104–6114.

(11) Manalo, M. N.; Prez, L. M.; Li Wang, A. *J. Am. Chem. Soc.* **2009**, *129*, 11298–11299.

(12) Tapavicza, E.; Lin, I.-C.; von Lilenfeld, A.; Tavernelli, I.; Coutinho-Neto, M. D.; Rothlisberger, U. *J. Chem. Theory Comput.* **2007**, *3*, 1673.

(13) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.

(14) Mignon, P.; Loverix, S.; De Proft, F.; Geerlings, P. *J. Phys. Chem. A* **2004**, *108*, 6038–6044.

(15) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

(16) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.

(17) Vazquez-Mayagoitia, A.; Sherrill, C. D.; Apra, E.; Sumpter, B. G. *J. Chem. Theory Comput.* **2010**, *6*, 727–734.

(18) Tsuzuki, S.; Lüthi, H. P. *J. Chem. Phys.* **2001**, *114*, 3949–3957.

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

(20) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* **1996**, *100*, 5590–5596.

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

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

(23) Mo, Y. *J. Mol. Model.* **2006**, *12*, 665–672.

(24) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334–338.

(25) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624–1626.

(26) Allen, M. J.; Tozer, D. J. *J. Chem. Phys.* **2002**, *117*, 11113–11120.

(27) Hobza, P.; Šponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

(28) Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

(29) Kurita, N.; Sekino, H. *Int. J. Quantum Chem.* **2003**, *91*, 355–362.

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

(31) Lee, T. J.; Scuseria, G. E. Achieving Chemical Accuracy with Coupled-Cluster Theory. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995.

(32) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford: New York, 1989; Volume *16* International Series of Monographs on Chemistry.

(33) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693–2699.

(34) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.

(35) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

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

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

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

(39) Chai, J.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(40) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(41) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(42) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

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

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

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

(46) Ushiyama, H.; Takatsuka, K. *J. Chem. Phys.* **2001**, *115*, 5903–5912.

(47) Shetty, S.; Pal, S.; Kanhere, D. G.; Goursot, A. *Los Alamos Natl. Lab., Prepr. Arch., Condens. Matter* **2004**, 1–30.

(48) Podolyan, Y.; Gorb, L.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 12103–12109.

(49) Miura, S.; Tuckerman, M. E.; Klein, M. L. *J. Chem. Phys.* **1998**, *109*, 5290–5299.

(50) Kim, Y.; Lim, S.; Kim, Y. *J. Phys. Chem. A* **1999**, *103*, 6632–6637.

(51) Dedikova, P.; Pitonak, M.; Neogrady, P.; Cernusak, I.; Urban, M. *J. Phys. Chem. A* **2008**, *112*, 7115–7123.

(52) Aplincourt, P.; Bureau, C.; Anthoine, J.-L.; Chong, D. P. *J. Phys. Chem. A* **2001**, *105*, 7364–7370.

(53) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157–9167.

(54) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.

(55) ACES II program is a product of the Quantum Theory Project, University of Florida. Authors: Stanton, J. F. Gauss, J. Watts, J. D. Nooijen, M. Oliphant, N. Perera, S. A. Szalay, P. G. Lauderdale, W. J. Gwaltney, S. R. Beck, S. Balková, A. Bernholdt, D. E. Baeck, K.-K. Sekino, H. Rozyczko, P. Huber, C. Bartlett. R. J. Integral packages included are VMOL (Almöf, J. Taylor, P. R.), VPROPS (P. R. Taylor), and a modified version of the ABACUS integral derivative package(Helgaker, T. U. Aa. Jensen, H. J. Olsen, J. Jørgensen, P. Taylor P. R.).

(56) Werner, H.-J. MOLPRO, version 2009.1. http://www.molpro.net (accessed Nov 2010).

(57) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

(58) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; Di Labio, G. A. *J. Chem. Phys.* **2009**, *131*, 034111.

(59) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(60) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(61) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.

(62) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.

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

(64) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(65) Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

(66) The DFT-D2 $s_6$ parameter for the B970 functional was obtained by minimizing the mean absolute percent deviation for the S22 test set calculated with un-counterpoise-corrected and equally weighted aug-cc-pVDZ and aug-cc-pVTZ basis sets. A more thorough procedure, taking into account BSSE-corrected interaction energies, shifts the recommended parameter only slightly, to 0.80.

(67) Burns, L. A.; Vázquez-Mayagoitia, Á.; Sumpter, B. G.; Sherrill, C. D. Manuscript in preparation.

(68) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(69) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

(70) Williams, H. L.; Szalewicz, K.; Jeziorski, B.; Moszynski, R.; Rybak, S. *J. Chem. Phys.* **1993**, *98*, 1279–1292.

(71) Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 184111.

(72) Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610–1616.

(73) Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *133*, 014101.

(74) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

(75) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.

CT100469B

JCTC Journal of Chemical Theory and Computation

# Nature of the Carbon−Sulfur Bond in the Species H−CS−OH

Henry S. Rzepa*

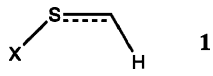
*Department of Chemistry, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom*

Received August 20, 2010

Ⓦ **This paper contains enhanced objects available on the Internet at http://pubs.acs.org/JCTC.**

**Abstract:** A QTAIM (Quantum-Theory-Atoms-in-Molecules) and ELF (electron localization function) topological analysis of the bonding in the recently reported molecule HCSX, X = OH reveals that the central carbon−sulfur bond is highly tunable, from having triple character at one limit to being almost a single bond at the other depending on the nature of the group X.

The nature of bonding, and in particular the degree of multiple bonding between particular pairs of atoms continues to attract much attention from experimentalists, theoreticians, and increasingly from groups which exploit the synergies between both. Such an example appeared recently documenting the preparation, spectroscopic characterization and theoretical analysis of the species H−CS−X (**1**, *X* = OH).[1] The focus was on the nature of the carbon−sulfur bond, and its multiple-bonded character. The conclusion drawn from various types of analysis of the computed wave function and comparison with measured stretching frequencies was that the species exhibits a "*rather strong C═S double bond, or a weak C≡S triple bond*".[1] In this communication, I suggest that the multiple character of this carbon−sulfur bond can in fact be uniquely tuned across an unusually wide range simply by variation of the group X in **1**.




**1**

The original analysis[1] was based, inter alia, on comparison of measured C−S stretching frequencies with force constants computed at the CCSD (T)/cc-pVTZ level. Analysis of the wave function was provided at this level by inspection of individual molecular orbitals and localized NBO analysis of the bond orders, and by comparison with the diatomic species CS itself. Two methods that were not included in the original discussion[1] were those based on the topological properties of the electron density (QTAIM)[2] and analysis of the electron localization function (ELF) $\eta$.[3,4] These methods can often provide illuminating and complementary insights into the nature of bonding,[5] and they formed the basis for a discussion of the original article[1] on Bachrach's blog.[6] The discussion continued on this author's own blog,[7] with a formal outcome which includes further computational analysis being presented here.

The QTAIM method identifies, inter alia, the coordinates of a particular type of saddle-point in the electron density distribution ($\rho(r)$, which can be either computed as here, or measured experimentally) and which is known as a bond critical point, henceforth referred to as a BCP. Three electronic properties at the BCP are commonly used to characterize the nature of the bonding in this region; the value of $\rho(r)$ itself, the bond ellipticity $\varepsilon$[8], and the Laplacian $\nabla^2\rho(r)$. The magnitude of $\rho(r)$ varies with the degree of multiple bonding, but it does have to be calibrated against other systems with the same atom constituents; in other words, it is a relative rather than absolute index. The bond ellipticity $\varepsilon$, which is a measure of the deviation of the electron density distribution from cylindrical symmetry, has also been used[8] in an empirical sense to distinguish between a double bond (for which $\varepsilon$ has typical values of 0.4−0.8) and a triple or single bond (where it is close to zero). The Laplacian $\nabla^2\rho(r)$ is a more interesting metric[9] related to the kinetic and potential energy densities at the BCP. Thus, negative values of $\nabla^2\rho(r)$ together with a high value for $\rho(r)$ are normally associated with a lowering of the potential energy and of covalent character in the bonding region. Positive values for $\nabla^2\rho(r)$ are indicative of excess kinetic energy density over potential energy density (i.e., $2G(r) > |V(r)|$, where $G(r)$ is the kinetic energy density and $V(r)$ the potential energy

* Corresponding author e-mail: rzepa@imperial.ac.uk.

10.1021/ct100470g 
Published on Web 12/09/2010

***Table 1.*** Calculated Properties for 1, CS, and $HCS^+$ [a]

| system | C−S length (Å), $\nu_{CS}(cm^{-1})$ [b] | AIM: $\rho(r)_{C-S}$; $\nabla^2\rho(r)$, $\varepsilon$ [c] | ELF, C−S basin integral (electrons) [d] | NBO E2 (kcal/mol), Wiberg bond order [e] | digital repository [f] |
|---|---|---|---|---|---|
| HC≡S$^+$ | 1.4902, 1403 | 0.275; +0.723, 0.00 | 3.53 [g] | -, 2.94 | 10042/to-3617<br>10042/to-3661 |
| 1, $X$ = OTf | 1.4898 [h],1393 | 0.283; +0.618, 0.022 [h] | 2.60 | -, -, --, 2.67 | 10042/to-3628<br>10042/to-5102 |
| C=S | 1.5505, 1269 | 0.276; +0.226, 0.00 | 2.73 | -, 2.70 | 10042/to-3616<br>10042/to-5107 |
| 1, $X$ = F | 1.5273, 1272 | 0.278; +0.150, 0.261 | 2.62 | -, 2.44 | 10042/to-3624<br>10042/to-3654 |
| 1, $X$ = Cl | 1.5352,1244 | 0.272;+0.163, 0.196 | 2.68 | 119.3,2.38 | 10042/to-3622<br>10042/to-3655 |
| 1, $X$ = HO | 1.5549, 1205 | 0.272; −0.114, 0.380 | 2.62 | 44.0, 2.27 | 10042/to-3627<br>10042/to-3653 |
| 1, $X$ = $H_2N$ | 1.5889, 1010/1131 | 0.267; −0.460, 0.459 | 2.39 | 24.5, 2.05 | 10042/to-5199<br>10042/to-5200 |
| 1, $X$ = CN | 1.6322, 842/1040 | 0.250; −0.575, 0.126 | 1.94 | 17.5, 1.81 | 10042/to-3638<br>10042/to-5106 |
| 1, $X$ = $H_2B$ | 1.7110, 814 | 0.221; −0.436, 0.052 | 1.89 | 6.1, 1.54 | 10042/to-3625<br>10042/to-3662 |
| $CH_3SH$ | 1.8255, 719 | 0.178; −0.278, 0.09 | 1.60 | -, 1.04 | 10042/to-3615<br>10042/to-5105 |

[a] An interactive version of this table is available via the HTML version of this article. Display requires a Java-enabled Web browser, 3D models, and 2D diagrams being invoked by clicking on any of the labeled buttons or pull-down menus in the left-hand navigation window. The resulting Jmol display can also be controlled by a pull down menu produced by a right-mouse click in the right-hand viewing window, from which individual coordinate files can also be acquired. [b] CCSD(T)/cc-pVTZ, calculated using Gaussian 09, Rev A.02. [c] Calculated at the CCSD/cc-pVTZ level using natural orbitals with the programs AIMALL V. 9.10.17 (aim.tkgristmill.com/) and AIM2000 (www.aim2000.de/). BCPs as shown as small purple spheres, and ring critical points as larger yellow spheres. The value of $\rho(r)$ at the critical point is shown as an attached label. [d] ELF function $\eta$ calculated at the CCSD/cc-pVTZ level using CCSD(T)/cc-pVTZ geometries, with natural orbitals (density = cc keyword) and the program TopMod09 (ref 12). The basin integrations are shown as attached labels, with the basin centroids shown as small magenta spheres. [e] NBO interaction energy (kcal mol$^{-1}$) between LP(carbon) and BD. * (S-X), calculated using B3LYP/cc-pVTZ wave functions at CCSD(T)/cc-pVTZ geometries for the NBO population analysis and the Wiberg bond index of the C−S bond. [f] OAI-PMH compliant Digital repository identifier, resolved as e.g. http://dx.doi.org/10042/to-2494. [g] Using TopMod09, a single ELF basin for the CS bond is identified, having a centroid located along the bond, with the total basin integration shown. Using DGrid-4.5, the ELF function is identified as a torus, with multiple basin centroids distributed around the center line of the torus but resulting in the same total integration. [h] Geometry optimized at the B3LYP/cc-pVTZ level and the wave function calculated at the CCSD/cc-pVTZ level.

density) and indicate either ionic or charge-shift character.[10] Ionic bonds, which experience closed-shell Pauli (also known as overlap or exchange) repulsions are associated with the contraction of electrons toward an atom and a low value for $\rho(r)$ at the BCP. In contrast, charge-shift bonds retain high values for $\rho(r)$, and bond stabilization for this latter type of bond is thought to originate from resonance terms between charge-shifted or ionic valence-bond forms and the (potentially repulsive) covalent form.[10]

A related approach is adopted for the (closed shell) ELF technique.[3,4,11,12] The wave function is used to compute an electron localization function $\eta$, which is normalized between values of 0.0 and 1.0.[3] This empirical function is based on comparing the additional kinetic energy density due to the Pauli principle at any point with that of a homogeneous electron gas (for which $\eta = 0.5$). By contouring the function at different isosurface thresholds, one can systematically locate the domains in $\eta$, and these can then be associated with either core electrons (which are referred to as attractors) or the valence electrons. The diatomic valence bonding regions are referred to as disynaptic basins (trisynaptic basins can also exist, and correspond to three-center bonds) and the nonbonding ones as monosynaptic basins. These basins can then be integrated for $\rho(r)$ to give an estimate of the electron population for each bond or lone pair, and from that an estimate of the bond order.[11]

The level of theory adopted here is the CCSD(T)/cc-pVTZ coupled cluster approach used previously,[1] all geometries being fully optimized at this level using Gaussian 09.[13] Full computational details and results are available via the digital repository entries listed in the Table 1. The AIM properties and ELF function were computed at the CCSD level using natural orbitals. The original ELF approach was reformulated for Kohn−Sham wave functions by Savin et al.[14] and more recently Silvi and co-workers have presented an extension to correlated and/or multireference methods.[12] This latter approach, implemented in the TopMod09 program has been used here, employing the CCSD method. The results are displayed in the online interactive Table 1 for a variety of different substituents X for species **1**, and also for three calibrants, $CH_3SH$, CS, and its protonated form HC≡S$^+$. The reason for including the latter is that in the limiting case where X is highly electron withdrawing (a good nucleofuge), species **1** dissociates to $HCS^+$ and $X^-$.

$HCS^+$ itself, being at one extreme of the series investigated (the other extreme being $CH_3SH$) in this analysis reveals itself to have something close to a C≡S triple bond. The bond length for this molecule is the shortest by a significant margin, the Wiberg bond order is almost 3, and the C≡S vibration mode at 1403 cm$^{-1}$ is easily the highest (it had been previously demonstrated[1] that the CCSD(T)/cc-pVTZ method matches experiment very well indeed in this regard). Of the AIM indices, $\rho(r)$ has a relatively high value of 0.275 au and $\varepsilon$ is zero, as it must be for a linear triple bond. For comparison, $\rho(r)$ for ethyne calculated at the same level is 0.418 au. The triflate **1**, X = OTf, in which the S−O bond

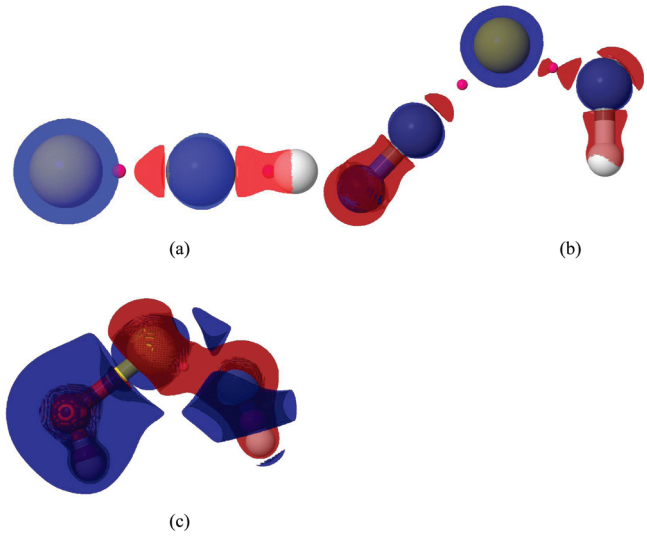



***Figure 1.*** Isosurface computed at CCSD/cc-pVTZ for (a) the computed Laplacian $\nabla^2\rho(r)$ for $HCS^+$ contoured at +0.723 (blue surface) and −0.723 (red surface). (b) $\nabla^2\rho(r)$ for **1**, X = CN contoured at +0.575 (blue surface) and −0.575 (red surface). (c) $\nabla^2\rho(r)$ for **1**, *X* = OH. Purple spheres are again indicative of the positions of BCPs.

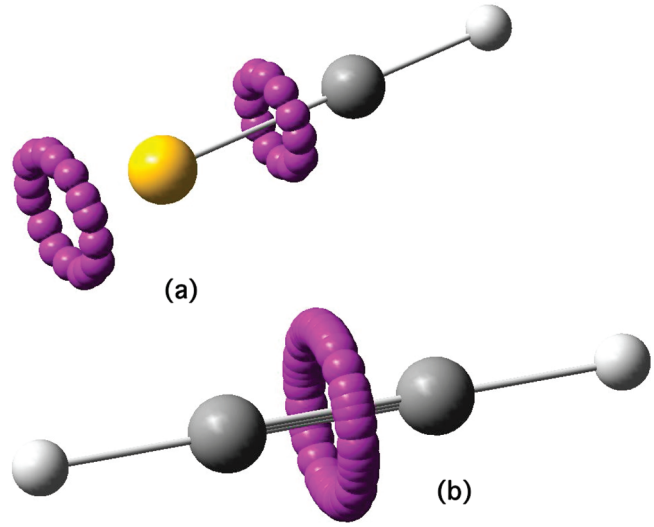



***Figure 2.*** Analysis of the ELF function for (a) $HCS^+$ and (b) HCCH. The centroids of the ELF localization domains (purple spheres) trace out the center line of torus encircling the CS or CC bond and (for $HCS^+$) the terminal lone pairs. The finite number of centroids seen in the diagram is purely a function of the resolution chosen for the cube of ELF values.

is largely ionic, resembles $HCS^+$ fairly closely (Table 1). The substituents **1**, X = F, Cl are however significantly attenuated from the triflate in $\nabla^2\rho(r)$ and **1**, X = OH (the species reported by Schreiner and co-workers[1]) continues this trend. Even more extreme attenuation can be achieved by using, e.g., the electropositive substituent **1**, X = $BH_2$. Here it is apparent that the C−S bond is only slightly stronger that a pure C−S single bond as exhibited by $CH_3SH$. The C−S bond ellipticity $\varepsilon$ reaches a maximum in the series for **1**, X = $H_2N$ (0.47); X = OH is somewhat reduced (0.38) indicating possibly the incursion of some triple bond character with this substituent (thus the description originally proposed[1] of it having a weak C≡S triple bond seems justified).

A large and positive value for the Laplacian $\nabla^2\rho(r)$ is calculated for $HCS^+$ (+0.723); this combination of large values for both $\rho(r)$ and $\nabla^2\rho(r)$ has been suggested as a characteristic of the charge-shift bond type recently highlighted by Shaik and Hiberty.[10] The Laplacian $\nabla^2\rho(r)$ across the entire series changes from positive (+0.618 for **1**, *X* = TfO) to negative (−0.575 for **1**, X = CN) implying a dramatic change in character of this bond from charge-shift to covalent induced by a mere neutral substituent X! The value of $\nabla^2\rho(r)$ for **1**, *X* = OH (−0.114) is at the crossover between the C−S bond having charge-shift and having covalent character. However, one must remember that the association between positive values of the Laplacian $\nabla^2\rho(r)$ and the valence-bond-computed bond stabilization deriving from resonance terms between charge-shifted VB structures and (more repulsive) covalent VB forms, as described by Hiberty and Shaik,[10] was largely derived from studies of homonuclear bonds from the first row. Less is known about such associations for second period elements such as sulfur.

To illustrate how care has to be taken in interpreting $\nabla^2\rho(r)$, the calculated isosurface for this property contoured at $\nabla^2\rho(r)$ = 0.723 (the value at the coordinate of the C−S BCP) is shown in Figure 1a for $HCS^+$ (an interactive version of this and other surfaces and coordinates is available via Table 1 in the HTML version of this article). The blue isosurface in the figure represents the positive Laplacian and is conventionally associated with (electrophilic) regions of charge depletion; the red surface represents the negative Laplacian and is associated with (nucleophilic) regions of charge accumulation. Typically, charge depletion occurs from the valence shells of atoms and the accumulation is found in the bonds. The BCP for the C−S bond of $HCS^+$ is typically displaced along the C−S axis away from the midpoint and toward the more diffuse distribution of the second period sulfur atom (Figure 1, see centroids of the purple spheres), where it encounters the charge-depleted region of the S valence shell. The C−S bond nonetheless does have a (red) region of charge accumulation, but this resides much closer to the carbon atom and does not envelope the bond-critical point. In contrast, the bond-critical coordinate of the C−H bond is displaced toward the hydrogen (reflecting the much smaller density around H and because as a general rule the BCP tends to be closer to the more electropositive atom of the bond), where it is contained within the region of negative (red) Laplacian.

A wider perspective can be obtained by comparing these results for $HCS^+$ with other systems. Discussing first **1**, *X* = CN, the calculated Laplacian $\nabla^2\rho(r)$ has a value (−0.575) for the C−S BCP which is at the negative extreme of the systems studied (Figure 1b). The electronic impact of substituent X has been to move the C−S BCP coordinate away from the sulfur, and toward the now more electropositive carbon. At this position, the BCP is now contained within the region of C−S bond charge accumulation, as suggested by the negative sign of the Laplacian at that point. The next system **1**, X = OH shows a value for $\nabla^2\rho(r)$ at the C−S BCP that is numerically small (+0.114, Figure 1c) and which reveals more complex features for the $\nabla^2\rho(r)$ isosur-

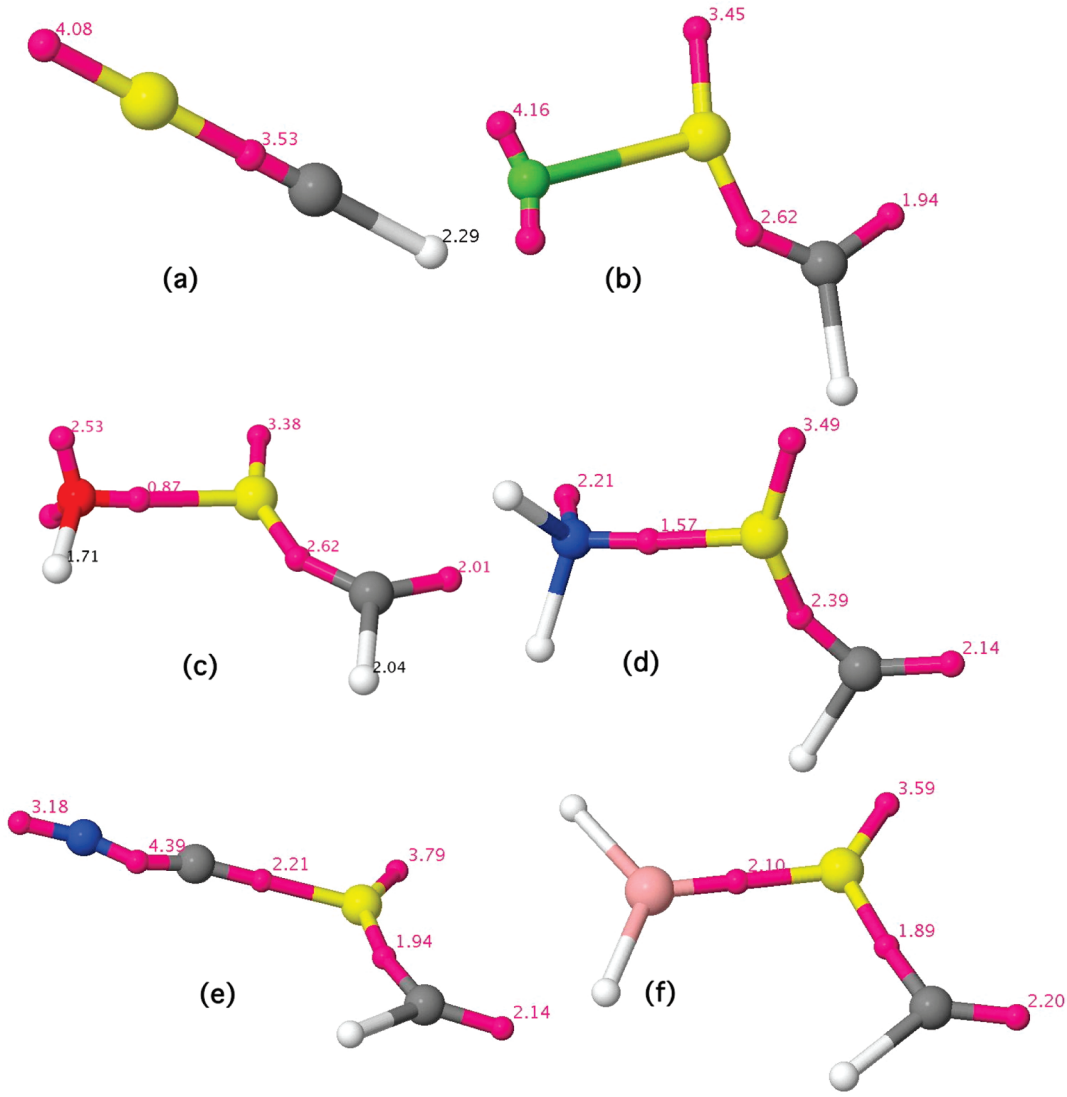



***Figure 3.*** ELF basin centroids (purple spheres) and basin electron populations calculated at the CCSD/cc-pVTZ level for (a) $HCS^+$, (b) **1**,X = F, (c) **1**,X = OH, (d) **1**,X = $NH_2$, (e) **1**,X = CN, and (f) **1**,X = $BH_2$.

face. The BCP is now located in a region where $\nabla^2\rho(r)$ is itself changing sign (in effect $\nabla^3\rho(r)$ is large) but the C−S bond itself is wrapped in a (red) torus of negative charge accumulation that spans both the bond and the adjacent lone pairs on C and S. The final discussion in this perspective relates to the pair **1**, X = $NH_2$ and X = $BH_2$, which (at first sight surprisingly) present rather similar $\nabla^2\rho(r)$ values at the C−S BCP (Table 1), despite the very different nature of the two substituents. This is due to a combination of the shape of the $\nabla^2\rho(r)$ isosurface and the actual position of the BCP. For X = $BH_2$, the BCP is more or less at the midpoint of the C−S bond, displaced very slightly toward the C, whereas with X = $NH_2$ the BCP is markedly displaced away from the midpoint toward the S. This only serves to indicate that the value of $\nabla^2\rho(r)$ at the BCP itself may not represent the character of the bond as a whole.

One may conclude from these trends that, unlike (homonuclear) bonds between first period elements, the sign of the Laplacian at the BCP for bonds involving a combination of first and second period element may not necessarily represent a simple measure of the degree of covalency or charge-shift character in that bond.

The ELF analysis of these systems adds several further insights to the bonding. The form of the ELF valence localization domains for the C≡S and sulfur lone pair regions of the axially symmetric species $HCS^+$ take the form of a **torus** encircling the CS bond (Figure 2) with a radius of ∼0.33 Å and second torus of radius 0.45 Å representing the S lone pairs. Ethyne itself[4] gives an ELF torus with radius 0.48 Å (Figure 2b) for the C≡C region which integrates to a population of 5.11 electrons. The total is less than the nominal six electrons for a triple bond because some absorption takes place into the carbon core attractors (∼0.1 each) and the C−H basins (∼0.35 each). Integration of the ELF torus encircling the CS bond in $HCS^+$ for $\rho$(r) yields a population of 3.53e, reduced from the nominal triple-bond population of six due to relocation of ∼0.3e into the CH basin, and ∼1.87e into the terminal sulfur lone pair region (Figure 2a). A similar reduction is found for, e.g., $N_2$, for which the N≡N basin population is reduced to ∼3.57e by transfer occurring from the $\pi$ region to the terminal lone pairs (3.11e each).

If the axial symmetry is destroyed, then the ELF function avoids the toroidal form of the localization domain, and

instead collapses into two or more[4] conventional disynaptic basins, depending on the symmetry. For the systems **1**, the perturbation by substituent X causes the carbon−sulfur localization domain to bifurcate into two such separate basins, one being disynaptic and localized into the axis of the CS bond (Figure 3). A second (formally disynaptic) basin is more biased toward the carbon and therefore tends to being a monosynaptic carbene "lone pair". Several associated trends are worth highlighting:

1. The population of the X−S basin is the most variable, ranging from zero (a significantly ionic bond, the covalent population being absorbed into lone pairs on X), through 0.87 (semi-ionic, X = OH), to ∼2.1−2.2 (covalent, X = $BH_2$, CN).

2. The population of the disynaptic C−S basin modestly decreases with decreasing electronegativity of X, from 2.62 (X = F) to 1.89 (X = $BH_2$).

3. The population of the second "carbene" basin remains relatively constant (1.94−2.20).

4. The sum of the C−S and "carbene" populations (4.63 for X = OH) is rather greater than that of the original unbifurcated triple bond C−S basin in $HCS^+$ itself (3.53). This is mostly due to scavenging of the sulfur lone pair population (4.08 for $HCS^+$ is reduced to 3.38 for X = OH).

5. Groups such as X = CN have a more complex interplay between the $\sigma$ and $\pi$ frameworks, overall tending to an electropositive manifestation for X.

This bifurcation (Figure 3) also now offers an opportunity to provide some insight into why this CS bond is so tunable across such a wide range (Table 1), from close to a triple bond (X = OTf) to being a single bond with X = $BH_2$. An NBO analysis[15] (of the DFT density) reveals the carbene basin ("Lp" in NBO terminology) to have a hugely variable interaction energy (Table 1) with the S−X BD* antibonding acceptor. The geometric alignment between the axis connecting the C to the centroid of the carbon basin, and the axis of the S−X bond is almost perfectly antiperiplanar; a classic anomeric effect in fact! One might also note another smaller anomeric alignment between the C−H bond as acceptor and the monosynaptic basin on the sulfur as donor (E2 interaction energy 7.8 kcal $mol^{-1}$). These interactions were not identified for comment in the original NBO analysis.[1]

For some substituents such as **1**,X = OTf or F, the S−X bond is so ionic that this term is not presented in the NBO analysis[1] (another manifestation of the absorption of the ELF disynaptic basin away from the S−X region for these molecules onto the monosynaptic oxy-anion lone pairs of X, as noted earlier). In the other direction, the disynaptic S−X bond basin integration reaches a maximum of 2.10e for **1**,X = $BH_2$, with a concomitant reduction to 1.89e for the C−S bond. The actual molecule **1**,X = OH that provoked this entire discussion is intermediate between these extremes; the S−X basin has a small population 0.87e (typical of a partially ionic bond) and the C−S basin integrates to 2.62e (some way off the maximum of 3.53e for the CS toroidal localization domain achieved by the fully ionised $HCS^+$ itself). The reduction in the CS basin integral is due to the bifurcation of the erstwhile triple bond basin into 2.01e that have split off to form a carbon "lone pair", leaving 2.62e behind in the CS region proper. This former density is perfectly oriented to achieve stabilization by an anomeric-style interaction with the S−X (antiperiplanar) acceptor bond.

This mechanism thus reveals the C−S "bond" in such systems to comprise two separate electronic components as identified by the ELF technique, the partitioning of which enables the high degree of tunability of the bond order. The QTAIM properties, and in particular the Laplacian at the C−S bond critical-point, also show considerable variation with the nature of group X. A full valence-bond analysis of these systems would be needed to establish if the C−S bond exhibits charge-shift character. A search identifying other molecules containing such tunable bonds may be a worthwhile endeavor.

**Acknowledgment.** The author thanks Philippe Hiberty, Sason Shaik, and Bernard Silvi for many helpful discussions.

## References

(1) Schreiner, P. R.; Reisenauer, H. P.; Romanski, J.; Mloston, G. *Angew. Chem.* **2009**, *48*, 8133–8136, DOI: 10.1002/anie.200903969.

(2) (a) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, UK, 1990; (b) Popelier, P. L. A., *Atoms in Molecules: An Introduction*; Prentice-Hall: London (2000); (c) Poater, J.; Duran, M.; Sola, M.; Silvi, B. *Chem. Rev.* **2005**, *105*, 3911–3947. (d) Bader, R. F. W. *J. Phys. Chem.* **2010**, *114*, 7431–7444, DOI: 10.1021/jp102748b.

(3) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–5403, DOI: 10.1063/1.458517.

(4) (a) For a wide ranging early review of the applications of this technique, see: Savin, A.; Nesper, R.; Wengert, S.; Fassler, T. E. *Angew. Chem., Int. Ed.* **1997**, *36*, 1808–1832, DOI: 10.1002/anie.19971808. For representative recent applications, see: (b) Berski, S.; Latajka, Z.; Gordon, A. J. *J. Chem. Phys.* **2010**, *133*, DOI: 10.1063/1.3460593. (c) Brock, D. S.; de Pury, J. J. C.; Mercier, H. P. A.; Schrobilgen, G. J.; Silvi, B. *J. Am. Chem. Soc.* **2010**, *132*, 3533–3542, DOI: 10.1021/ja9098559. (c) Contreras-Garcia, J.; Pendas, A. M.; Recio, J. M.; Silvi, B. *J. Chem. Theory Comput.* **2009**, *5*, 164–173, DOI: 10.1021/ct800420n. (d) Gotz, K.; Kaupp, M.; Braunschweig, H.; Stalke, D. *Chem.—Eur. J.* **2009**, *15*, 623–632, DOI: 10.1002/chem.200801073. (e) Grabowsky, S.; Hesse, M. F.; Paulmann, C.; Luger, P.; Beckmann, J. *Inorg. Chem.* **2009**, *48*, 4384–4393, DOI: 10.1021/ic900074r. (f) Jubert, A.; Okulik, N.; Michelini, M. D.; Mota, C. J. A. *J. Phys. Chem. A* **2008**, *112*, 11468–11480, DOI: 10.1021/jp805699x. (g) Lein, M. *Coord. Chem. Rev.* **2009**, *253*, 625–634, DOI: 10.1016/j.ccr.2008.07.007. (h) Matito, E.; Sola, M. *Coord. Chem. Rev.* **2009**, *253*, 647–665, DOI: 10.1016/j.ccr.2008.10.003. (i) Trujillo, C.; Mo, O.; Yanez, M.; Silvi, B. *J. Chem. Theory Comput.* **2008**, *4*, 1593–1599, DOI: 10.1021/ct800178x. (j) Vidal, I.; Melchor, S.; Dobado, J. A. *J. Phys. Chem. A* **2008**, *112*, 34143423 DOI: 10.1021/jp075370p. (k) Polo, V.; Andres, J.; Silvi, B. *J. Comput. Chem.* **2007**, *28*, 857–864, DOI: 10.1002/jcc.20615. (l) Mo, O.; Yanez, M.; Pendas, A. M.; Del Bene, J. E.; Alkorta, I.; Elguero, *J. Phys. Chem. Chem. Phys.* **2007**, *9*, 3970–3977, DOI: 10.1039/B702480K.

(5) Rzepa, H. S. *Nat. Chem.* **2009**, *1*, 510–512, DOI: 10.1038/nchem.373.

(6) Bachrach, S. The C-S triple bond. URL:http://comporgchem.com/blog/?p=510. Accessed: 2010−04−20. (Archived by WebCite® at http://www.webcitation.org/5p8JYMMba).

(7) (a) Rzepa. H. S. The nature of the C'S triple bond. URL: http://www.ch.ic.ac.uk/rzepa/blog/?p=1210. Accessed: 2010−04−20. (Archived by WebCite® at http://www.webcitation.org/5p8Jp1zNh); (b) Rzepa, H. S. The nature of the C'S Triple bond: Part 2. URL:http://www.ch.ic.ac.uk/rzepa/blog/?p=1243. Accessed: 2010−04−20. (Archived by WebCite at http://www.webcitation.org/5p8KGaf5n); (c) Rzepa, H. S. The nature of the C'S triple bond: part 3. URL:http://www.ch.ic.ac.uk/rzepa/blog/?p=1278. Accessed: 2010−04−20. (Archived by WebCite at http://www.webcitation.org/5p8KMlrLF).

(8) For a succinct definition and discussion of these properties, see Rode, J. E.; Dobrowolski, *J. Chem. Phys. Lett* **2007**, *449*, 240–245, DOI: 10.1016/j.cplett.2007.10.048.

(9) For a useful definition and discussion of this property, see: Kobayashi, K.; Nagaser, S. *Chem. Phys. Lett.* **1999**, *302*, 312–316, DOI: 10.1016/S0009−2614(99)00135−9.

(10) (a) Shaik, S.; Danovich, D.; Wu, W.; Hiberty, P. C. *Nature Chem.* **2009**, *1*, 443–449, DOI: 10.1038/nchem.327. (b) Shaik, S.; Danovich, D.; Silvi, B.; Lauvergnat, D. L.; Hiberty, P. C. *Chem.—Eur. J.* **2005**, *11*, 6358–6371, DOI: 10.1002/chem.200500265.

(11) (a) Savin, A. *J. Chem. Sci.* **2005**, *117*, 473–475, DOI: 10.1007/BF02708351. (b) Chamorro, E.; Fuentealba, P.; Savin, A. *J. Comput. Chem.* **2003**, *24*, 496–504, DOI: 10.1002/jcc.10242.

(12) Feixas, F.; Matito, E.; Duran, S. M.; Silvi, B. *J. Chem. Theory Comp.* **2010**, *6*, 2736−2742.

(13) *Gaussian 09, Revision A.2*, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, N. J., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., Fox, D. J. Gaussian, Inc.: Wallingford CT, 2009.

(14) (a) Savin, A.; Jepsen, O.; Flad, J.; Andersen, O. K.; Preuss, H.; von Schnering, H. G. *Angew. Chem. Int. Ed.* **1992**, *31*, 187–188, DOI: 10.1002/anie.199201871. (b) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683–686, DOI: 10.1038/371683a0. (c) Kohout, M.; Savin, A. *J. Comput. Chem.* **1998**, *18*, 1431–1439, DOI: 10.1002/(SICI)1096−987X(199709)18:12 < 1431::AID-JCC1 > 3.0.CO;2-K.

(15) Weinhold, F. Landis, C. R. *Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective*; Cambridge University Press: New York, 2005, pp 760.

CT100470G

JCTC Journal of Chemical Theory and Computation

# Valence Excited States in Large Molecules via Local Multireference Singles and Doubles Configuration Interaction

Tsz S. Chwee[†] and Emily A. Carter*[,‡]

*Departments of Chemistry and Mechanical and Aerospace Engineering and the Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544-5263, United States*

Received August 25, 2010

**Abstract:** We demonstrate that valence excited states in large molecules can be treated using local multireference singles and doubles configuration interaction (LMRSDCI). The interior eigenvalues corresponding to the excited states of interest are transformed and shifted to the extrema of the spectrum by way of oblique projections and a matrix shift within a modified Davidson diagonalization scheme. In this way, the approximate wave function associated with the excited state of interest can be isolated independently of the lower lying roots, and residual minimization is used for final convergence to the target eigenstate. We find that vertical excitation energies calculated using LMRSDCI are mostly within 0.2 eV of nonlocal MRSDCI values.

## 1. Introduction

The characterization of molecules in their electronically excited states finds relevance in many areas and applications ranging from the study of photophysical and photochemical processes in molecules to the development of solar cells in material science. For instance, there is considerable interest in the photophysical behavior of nucleic acid bases such as cytosine and thymine when exposed to ultraviolet radiation, as the resulting DNA damage may have cytotoxic/genotoxic consequences.[1,2] Separately, insight into chemical dynamics may rely on characterization of optically dark states that are inaccessible to photoexcitation/photoabsorption studies. While complementary techniques such as electron impact methods may be used to probe such states, they are impeded by inherently low resolution. As such, theoretical methods with capabilities beyond the characterization of the electronic ground state are important tools in aiding the understanding of the above-mentioned phenomena.

Indeed, many of the *ab initio* electronic structure methods commonly used to study molecules in the electronic ground state can be adapted for excited states. For example, in the method of configuration interaction singles (CIS), the CI Hamiltonian is diagonalized in the basis of singly excited determinants, and the higher roots serve as approximations to the excited states. Due to the lack of electron correlation, the quality of CIS calculations for the excited states is similar to that of Hartree−Fock (HF) for the ground state. Nevertheless, it has formed the basis for further improvements such as the inclusion of a perturbative doubles correction (CIS(D)),[3,4] spin-flip CIS,[5] and different variants of the XCIS[6,7] implementation.

Among methods that incorporate the effects of electron correlation, time-dependent DFT (TDDFT)[8,9] is one of the most computationally affordable. The foundation for extending conventional, time-independent DFT into the time domain is based on the Runge−Gross theorem,[8] which states the existence of a unique mapping between the time-dependent external potential and the density of a system. The relatively low cost of TDDFT computations facilitates the study of larger molecules than possible with other approaches, although there are well documented limitations with using this approach to study (i) Rydberg-type excitations, (ii) multielectron excitations, and (iii) excitations involving charge transfer.[10−13] Nevertheless, there has been progress in describing long-range charge-transfer excitations within DFT by imposing constraints (e.g., charge differences between the donor and acceptor fragments) within the ground

* Corresponding author e-mail: eac@princeton.edu.

[†] Department of Chemistry.

[‡] Department of Mechanical and Aerospace Engineering and the Program in Applied and Computational Mathematics.

10.1021/ct100486q 
Published on Web 12/03/2010

state formalism. The dominant Coulombic interaction between the charge transfer fragments can be correctly modeled using this approach.[14−16]

Within the domain of single-determinant-based wave function approaches, methods based on coupled cluster theory (CC) have been used extensively for studying excited states in small to medium sized molecules. Provided that the HF wave function serves as a good reference for the ground state and the excited states are single configurational in character, low order CC-based approaches such as equation-of-motion CC (EOM-CC)[17,18] and the closely related linear-response CC (LR-CC)[19−21] are among the most accurate approaches for treating singly excited states, yielding excitation energies within 0.3 eV of experimental values. While full EOM-CC and LR-CC are formally suitable for treating all excited states, current practical implementations of EOM-CC and LR-CC include up to the triple excitation cluster operator, i.e., CCSDT,[22,23] which scales as $O(N^8)$; higher order formulations such as CCSDTQ[24,25] are too expensive apart from benchmarking purposes. To reduce the cost associated with the family of CC methods, various approximations have been introduced. For example, in the method of CC2,[26] all terms in the $\hat{T}_2$ equations higher than second-order are neglected, reducing the overall computational cost to $O(N^5)$ from $O(N^6)$ in CCSD.[27] CC2 excitation energies are only slightly inferior to those of CCSD. However, the performance of low order CC theories degrades quickly when the ground state deviates from a single reference description. Thus, these theories should be used only for spectroscopic predictions near the equilibrium geometry of molecules well described by a single reference. The inclusion of higher order excitations, such as CCSDT (or the cheaper CC3),[28] can help to recover some static correlation. Nevertheless, the accuracy of such single reference methods is further compromised if the excited states are multiconfigurational in nature. Taking the example of $C_2$, there remains a deviation of 0.4 and 0.85 eV in the calculated excitation energies from full configuration interaction results at the CCSDT and CC3 levels of theory, respectively.[29,30] The spin-flip approach introduced by Krylov[31,32] provides one route forward within CC theory to deal with molecules that have near degeneracies. In brief, the model employs a high spin reference state that is well-represented by a single determinant while the target states are described as spin-flipping excitations from the reference state. The success of this model has been attributed to a balanced treatment of the targeted states.

Nevertheless, a multiconfigurational starting point for inclusion of electron correlation is still required in situations where the electronic structure is complex, e.g., in atoms with open shells such as transition metals, molecules with weakly coupled electrons such as diradicals, and photodissociation. The treatment of these quasi-degenerate effects (static correlation) can be handled using the complete active space self-consistent field (CASSCF)[33,34] or the less costly restricted active space self-consistent field (RASSCF)[35] methods to obtain a qualitatively adequate description of the wave function. Subsequently, the dynamic electron correlation component can be recovered using second-order perturbation theory (CASPT2[36,37] or RASPT2[38]), and calculations performed with these methods ($O(N^5)$ scaling) generally yield excitation energies within 0.2 eV of experimental values for organic molecules.[39,40] The method has also enjoyed success in inorganic spectroscopic studies involving transition metals.[41−43] Alternatively, multireference singles and doubles configuration interaction (MRSDCI) may be used to treat dynamic electron correlation effects. The reference configurations may be taken entirely from the preceding CASSCF calculation or individually selected on the basis of their contribution to the overall wave function. Just as the lowest root corresponds to the electronic ground state in the diagonalization of the CI Hamiltonian matrix, so the higher roots are associated with the electronically excited states. The accuracy of this approach has been demonstrated for small- and medium-sized molecules[44] even though a significant drawback of the method lies in its conventionally high computational cost ($O(N^6)$).

To lower the overall cost and extend the approach to larger molecules, local correlation has been implemented within several electronic structure methods such as EOM-CCSD[45,46] and CC2[47−50] for studying excited states and molecular properties. For electronic excitations that are localized in nature (i.e., not charge-transfer type excitations), it was found to be advantageous to define an excitation domain[45] that may be predetermined via a less costly calculation, e.g., CIS. In addition, double excitations from localized occupied orbitals are segregated into strong and weak pairs, and computational savings may be derived by restricting explicit correlation of electrons to those within the strong pairs. In both the local EOM-CCSD and CC2 implementations, the calculated excitation energies for a range of medium-sized organic molecules (such as propanamide and tyrosine) were found to differ from the nonlocal calculations by less than 0.2 eV.[45−47]

Thus far, MRSDCI algorithms have been constructed within a local correlation framework (LMRSDCI) only for ground electronic states.[51−56] Here, we extend our LMRSDCI method to the treatment of valence excited states. A modified Davidson diagonalization (DD) scheme[57−60] is used in a preliminary step to obtain a refined guess for the wave function of the targeted excited state without explicitly solving for the lower roots. Once an approximate wave function is available, residual minimization via the residual minimization method−direct inversion in iterative subspace (RMM-DIIS)[61−66] scheme is used for final convergence. Using this protocol, we verify its efficacy for vertical excitation energies within LMRSDCI for various molecules (containing up to nine heavy atoms). In most cases, agreement to within 0.2 eV of nonlocal MRSDCI values is obtained. After completing this verification for the size of molecules for which nonlocal MRSDCI excitation energies are still calculable, we then use our LMRSDCI theory to predict excitation energies in some large molecules for which conventional MRSDCI could not be carried out.

## 2. Theory

**2.1. Local Correlation within Configuration Interaction.** Configuration interaction is one of the earliest methods used to account for the effects of electron correlation. The variational electronic wave function is written as a linear combination of Slater determinants (or spin-adapted configuration state functions (CSFs)) constructed from molecular orbitals.

$$\Psi^{\mathrm{CI}} = \sum_{R} c_R \Psi_R + \sum_{i,a} c_i^a \Psi_i^a + \sum_{ij,ab} c_{ij}^{ab} \Psi_{ij}^{ab} + \ldots \quad (1)$$

where $\{\Psi_R\}$ contains the set of reference configurations while $\{\Psi_i^a\}$ and $\{\Psi_{ij}^{ab}\}$ are singly and doubly excited configurations generated by promoting electrons in the reference configurations from internal orbitals $\{i,j\}$ to external orbitals $\{a,b\}$, respectively. A full CI calculation that includes all possible levels of excitations is not computationally feasible due to the prohibitive cost. As such, truncation of the CI basis is usually imposed at the level of double excitations since configurations generated from higher order excitations are not coupled directly to the reference wave function via the full Hamiltonian.[67,68] The length of the CI expansion in eq 1 may be reduced further by working within a local correlation framework that exploits the short-ranged nature of electron correlation. Using localized orbitals to span the occupied and virtual subspace, the correlated electron pairs are restricted to those in which the occupied orbitals $i$ and $j$ are spatially close.[51] Similarly, a correlation domain for some occupied orbital $i$ can be imposed to consist only of virtual orbitals that are localized in the vicinity of $i$.[51] Variational minimization of the CI energy with respect to the CI expansion coefficients in eq 1 leads to the usual eigenfunction−eigenvalue equation, $HC = EC$.

**2.2. Excited States within CI.** The lowest root of the eigenvalue problem corresponds to the molecular ground state, which can be isolated using iterative subspace projection methods like DD.[69] Information on the excited states can be similarly obtained from the higher lying roots, and adaptations of DD (e.g., the Davidson−Liu method[70]) are often used for this purpose. These implementations typically require the concurrent extraction of several lower lying roots or explicit orthogonalization to the lower roots, which in turn have to be found. Since diagonalization of the CI Hamiltonian matrix constitutes a major component of the overall computational cost, savings may be derived if the targeted eigenpair can be obtained in isolation. As an alternative, the targeted interior eigenvalue may be transformed and shifted to the extrema of the spectrum. The incorporation of such "shift-and-invert" methodology within DD may be accomplished by way of oblique projection techniques along with a matrix shift.[57−60] In the conventional implementation of DD that uses orthogonal projection, an approximate eigenvector of some large matrix $A$ is extracted from a lower dimensional subspace, $V \equiv \{v_1, v_2, \ldots, v_n\}$ (the search space). The approximate eigenvector, $\tilde{v}$, is written as a linear combination of the search space vectors, $\{v_i\}$, which constitute an orthonormal set.

$$\tilde{v} = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n \quad (2a)$$

$$A(Vc) = \tilde{\lambda}(Vc) \quad (2b)$$

where $c \equiv \{c_1, c_2, \ldots, c_n\}^T$, $\tilde{v} = Vc$, and $\tilde{\lambda}$ is the approximate eigenvalue.

The eigenvalue problem in eq 2b is underdetermined, but additional constraints are imposed via the Galerkin condition, which requires the residual $r = A\tilde{v} - \tilde{\lambda}\tilde{v}$ to be orthogonal to the test space. Within orthogonal projection, the test space is equivalent to the search space (i.e., $V$).

$$AVc - \tilde{\lambda}Vc \perp \{v_1, v_2, \ldots, v_n\} \quad (2c)$$

The orthogonality constraint leads to

$$V^*(A - \tilde{\lambda}I)Vc = 0 \quad (2d)$$

$$V^*AVc = \tilde{\lambda}V^*Vc = \tilde{\lambda}c \quad (2e)$$

On the other hand, the test space within oblique projection, $W$, is allowed to be different from the search space, $V$. In particular, by setting $W = AV$ and requiring that $W$ remain an orthonormal set,[57,59,60] the corresponding Galerkin−Petrov condition is

$$AVc - \tilde{\lambda}Vc \perp W, W \equiv \{w_1, w_2, \ldots, w_n\} \quad (3a)$$

$$W^*(A - \tilde{\lambda}I)Vc = 0 \quad (3b)$$

$$V^*A^*AVc = \tilde{\lambda}V^*A^*Vc = \tilde{\lambda}W^*Vc = \tilde{\lambda}W^*A^{-1}AVc = \tilde{\lambda}W^*A^{-1}Wc \quad (3c)$$

$$W^*A^{-1}Wc = \left(\frac{1}{\tilde{\lambda}}\right)c \quad (3d)$$

The low dimensional eigenvalue problem in eq 3d, transformed and reduced to the determination of eigenpairs belonging to $A^{-1}$ in the space of $W$, may be solved directly using conventional methods. $\tilde{\lambda}$ is the harmonic Ritz value,[57] and with an appropriate choice for the matrix shift, $\sigma$, $1/(\tilde{\lambda} - \sigma)$ can be transformed to the extremum of the spectrum. In our work, the modified DD scheme utilizing oblique projection along with a matrix shift is used as a preliminary step to generate a refined approximation for the wave function associated with the target eigenstate. Final convergence is achieved via residual minimization using the RMM-DIIS eigenvalue solver where the guess wave function obtained earlier now serves as the secondary input. DIIS was first developed by Pulay for accelerating SCF convergence.[61−63] Later, Wood and Zunger[64] and Hutter et al.[65] introduced RMM-DIIS for solving eigenvalue problems. Our implementation of RMM-DIIS follows the approach of Kresse.[66] To reiterate, the hybrid scheme involves two separate steps. First, a few iterations of a modified Davidson method are completed to arrive at a refined guess for the target eigenstate before it is input into the RMM-DIIS scheme for final convergence.

## 3. Computational Details

The ground state geometries of all molecules considered in this study were optimized within density functional theory using the hybrid B3LYP exchange-correlation functional[71] and the 6-31G** basis set. Our reported vertical excitation energies $T_v$ do not include zero point corrections, leading to some systematic errors in comparing to measured excitation energies $T_0$, since ground state vibrational frequencies are usually larger than excited state ones. However, the main purpose of our work is to test the accuracy of LMRSDCI for calculating $T_v$ compared to nonlocal MRSDCI and to demonstrate its efficacy in large molecules. When used in future applications, zero-point corrections can easily be accounted for at, e.g., the DFT-B3LYP or CASSCF level of theory.

Quasi-degeneracies are treated using state-averaged CASSCF, while dynamic electron correlation effects are taken into account by second order perturbation theory (CASPT2) and MRSDCI/LMRSDCI. These latter methods without local truncation are meant to provide benchmarks for LMRSDCI. All nonlocal calculations were carried out within the Molcas 7.2 quantum chemistry package.[72] CASSCF and CASPT2 results were obtained within Molcas using the RASSCF and CASPT2 modules, respectively. For the LMRSDCI calculations, inactive orbitals were localized using the Pipek−Mezey (PM) functional,[73] while the virtual space was spanned by localized orthogonal virtual orbitals generated using the scheme described by Subotnik et al.[74] To accelerate the convergence of the MRSDCI expansion as determined by a CI coefficient cutoff of 0.1 for the reference space, the active orbitals were segregated into strongly and weakly occupied (occupation number above 1.8 and below 0.2, respectively) and then separately localized using the PM functional within each set.[75] In addition, Cholesky vectors[76−78] were used in place of conventional two-electron integrals. The Cholesky vectors were generated via an incomplete Cholesky decomposition on the two-electron integral matrix using a decomposition threshold of $10^{-7}$.

The procedure for setting up orbital domains and other details of our LMRSDCI implementation can be found in our earlier work.[79−85] In brief, a sphere is associated with each molecular orbital in which a radius and a center of charge are determined. To demarcate the extent of each sphere, a list of atoms that contribute to the molecular orbital via the Mulliken population scheme is drawn up. The center of each sphere is calculated from the weighted average of the coordinates of the contributing atoms, while the radius is determined by taking the separation between the two atoms that contribute most heavily to the molecular orbital. A radius multiplier, which functions as a scaling parameter to the radius, is assigned to each sphere. The scaled value is the effective radius of the sphere and is necessary to allow for a systematic variation of the thresholds used within the weak pairs (WP) and truncation of virtuals (TOV) approximations in our local correlation studies. For example, under the WP approximation, two orbitals are deemed to be a weak pair if their associated spheres do not overlap, i.e., the separation between the spheres as given by the distance between the centers of the spheres is larger than the sum of their individual effective radii. Similarly, under the TOV approximation, the orbital domain of an internal orbital consists of a restricted set of virtual orbitals. The permitted virtual orbitals are those whose associated spheres overlap with the sphere of the internal orbital. In our work, the spheres associated with the internal (inactive and active) and external (virtual) orbitals residing on nonheteroatoms (i.e., carbon and hydrogen) were assigned a scaling factor of 1.5 and 1.2, respectively, while spheres associated with orbitals residing on the heteroatoms were assigned a larger scaling factor of 2.0.

For ground state MRSDCI/LMRSDCI calculations, the lowest roots were extracted using regular DD, and convergence to an energy difference less than $10^{-7}$ Hartrees between subsequent iterations was typically reached within 30 iterations. The higher lying roots corresponding to the excited states were isolated with the modified Davidson method described earlier. The parameter for the matrix shift was obtained from the corresponding eigenvalue of the preceding CASSCF calculation. After a refined initial guess was obtained from the modified Davidson scheme, the approximate CI vector was used as an input for the ensuing RMM-DIIS eigensolver. Convergence to a residual norm below $10^{-4}$ in between iterations required about 20−40 iterations.

## 4. Results

We first evaluate the performance of using the hybrid scheme described above within LMRSDCI for characterizing the valence excited states in various organic molecules for which some experimental data are available. We then go on to make some predictions for excited states in other large molecules.

We first consider the gas phase electronic transitions involving single excitations from the $\pi$ ($e_{1g}$, HOMO) to $\pi^*$ ($e_{2u}$, LUMO) orbitals of benzene ($D_{6h}$), which have been well studied experimentally. The possible excited states are determined from the direct product $e_{1g} \otimes e_{2u}$ giving rise to $^{1,3}B_{2u}$, $^{1,3}B_{1u}$, and $^{1,3}E_{1u}$ states. The singlet states are more important due to the spin selection rule $\Delta S = 0$ in electric-dipole transitions. In our work, we do not impose symmetry on the electronic wave function. Preliminary calculations on the ground $A_{1g}$, excited $B_{2u}$, and $E_{1u}$ states with $D_{2h}$ symmetry (the highest order point group available within Molcas) reveal that the vertical excitation energies differ from results obtained without imposing symmetry constraints on the wave function by less than 0.15 eV. The electronic term symbols used above are retained nevertheless, as they are useful for classification purposes.

For the preliminary state-averaged CASSCF calculation, we use a valence-type [6e,6o] active space comprising the six valence $\pi$ electrons distributed among the six $\pi$ and $\pi^*$ orbitals. For an explicit treatment of $\sigma-\pi$ correlation effects,[86−88] the inclusion of $\sigma$ electrons and the corresponding orbitals into the active space is necessary, i.e., a [12e,12o] active space. This is especially relevant for the so-called "ionic" $B_{1u}$ and $E_{1u}$ states, and therefore we expect some errors here due to the exclusion of $\sigma-\pi$ correlation. However, our goal here is to test the local correlation approximation for excited states and not to perform the most accurate

***Table 1.*** Vertical Excitation Energies $T_v$ (eV) from the $A_{1g}$ Ground State to the $B_{2u}$ Excited State of Benzene As Predicted by Various Methods and Basis Sets

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI | exptl $T_0$[89] |
|---|---|---|---|---|---|
| 6-31G* | 4.89 | 5.14 | 5.03 | 4.95 | 4.90 |
| 6-31G** | 4.88 | 5.13 | 5.05 | 4.96 | |
| 6-31+G* | 4.83 | 5.04 | 4.95 | 4.81 | |
| 6-31++G** | 4.82 | 5.02 | 4.95 | 4.82 | |
| 6-311G* | 4.87 | 5.10 | 5.05 | 4.95 | |
| 6-311+G* | 4.83 | 5.02 | 4.97 | 4.82 | |

***Table 2.*** Vertical Excitation Energies $T_v$ (eV) from the $A_{1g}$ Ground State to the $B_{1u}$ Excited State of Benzene As Predicted by Various Methods and Basis Sets

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI | exptl $T_0$[89] |
|---|---|---|---|---|---|
| 6-31G* | 8.10 | 6.80 | 6.90 | 7.08 | 6.20 |
| 6-31G** | 8.08 | 6.79 | 6.90 | 7.08 | |
| 6-31+G* | 7.96 | 6.40 | 6.45 | 6.67 | |
| 6-31++G** | 7.95 | 6.38 | 6.46 | 6.63 | |
| 6-311G* | 8.06 | 6.80 | 6.92 | 7.03 | |
| 6-311+G* | 7.92 | 6.39 | 6.45 | 6.61 | |

***Table 3.*** Vertical Excitation Energies $T_v$ (eV) from the $A_{1g}$ Ground State to the $E_{1u}$ Excited State of Benzene As Predicted by Various Methods and Basis Sets

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI | exptl $T_0$[89] |
|---|---|---|---|---|---|
| 6-31G* | 9.56 | 7.52 | 7.68 | 7.83 | 6.94 |
| 6-31G** | 9.54 | 7.47 | 7.61 | 7.82 | |
| 6-31+G* | 9.29 | 6.69 | 6.80 | 6.96 | |
| 6-31++G** | 9.27 | 6.70 | 6.83 | 7.00 | |
| 6-311G* | 9.41 | 7.45 | 7.64 | 7.80 | |
| 6-311+G* | 9.26 | 6.74 | 6.82 | 7.01 | |

possible CI calculations. The calculated vertical excitation energies relative to the $A_{1g}$ ground state are summarized in Tables 1−3.

For the lowest $B_{2u}$ excited state, the calculated vertical excitation energies are within 0.25 eV of experimental values for all levels of theory. The fact that CASSCF results are already close to experimental values is indicative of small differential correlation effects between the ground and excited states. However, this is not the case for the remaining "ionic" states of $B_{1u}$ and $E_{1u}$ symmetry where the CASSCF vertical excitation energies are off from experimental values by at least 1.7 eV. Incorporation of dynamic electron correlation within CASPT2 and MRSDCI improves the overall agreement with experimental values. In addition, a more extended basis may be useful in describing the diffuse character of some excited states. The improvements in the calculated vertical excitation energies for these "ionic" states are marked when diffuse functions are included on the carbon atoms, while similar augmentation on the hydrogen atoms has little influence on the final results. In all cases, LMRSDCI vertical excitation energies are within ∼0.2 eV of MRSDCI values.

***Table 4.*** Vertical Excitation Energies $T_v$ (eV) for the $n_O \rightarrow \pi^*$ Transition in Molecules 1−4 Shown in Figure 2

| basis/molecule | CASSCF | CASPT2 | MRSDCI | LMRSDCI |
|---|---|---|---|---|
| | | 6-31G* | | |
| 1 | 4.43 | 4.49 | 4.50 | 4.59 |
| 2 | 3.77 | 3.90 | 3.99 | 4.11 |
| 3 | 3.73 | 3.75 | 3.82 | 4.01 |
| 4 | 3.68 | 3.67 | 3.72 | 3.90 |
| | | 6-31+G* | | |
| 1 | 4.43 | 4.45 | 4.52 | 4.61 |
| 2 | 3.76 | 3.81 | 3.99 | 4.16 |
| 3 | 3.71 | 3.71 | 3.81 | 4.03 |
| 4 | 3.68 | 3.66 | 3.72 | 3.95 |
| | | 6-311G* | | |
| 1 | 4.42 | 4.47 | 4.52 | 4.61 |
| 2 | 3.74 | 3.86 | 4.01 | 4.12 |
| 3 | 3.71 | 3.71 | 3.80 | 3.95 |
| 4 | 3.65 | 3.61 | 3.74 | 3.93 |
| | | 6-311+G* | | |
| 1 | 4.39 | 4.44 | 4.55 | 4.66 |
| 2 | 3.69 | 3.80 | 4.02 | 4.11 |
| 3 | 3.67 | 3.66 | 3.84 | 4.01 |
| 4 | 3.63 | 3.57 | 3.72 | 3.95 |

In the following series of ketone molecules, 1−4 (Figure 1), the electronic transition arising from an excitation of an electron in one of the nonbonding lone pairs on the oxygen atom into the $\pi^*$ system ($n_O \rightarrow \pi^*$) is studied. For each molecule, the active space consists of the oxygen lone pair and the $\pi$ electrons distributed among the $\pi$ and $\pi^*$ system. Moving across the series from 1 to 4, there should be a concomitant decrease in the vertical excitation energies as the $\pi^*$ system is stabilized due to conjugation effects. This is captured in the calculated vertical excitation energies, which are listed in Table 4. In all cases, there is good agreement between CASPT2 and MRSDCI values while LMRSDCI results are within 0.25 eV of MRSDCI values.

The nature of excited states in nucleic acid bases, like cytosine, is of interest because these molecules are the

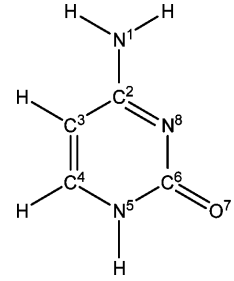



***Figure 2.*** Cytosine, a nucleic acid base.

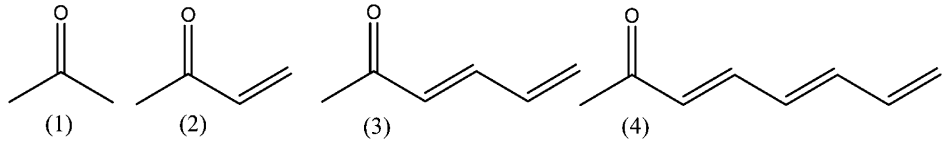



***Figure 1.*** Test set for the study of vertical excitation energies corresponding to the $n_O \rightarrow \pi^*$ transition. The experimentally determined vertical excitation energy $T_0$ for molecule 1 is 4.49 eV.[90]

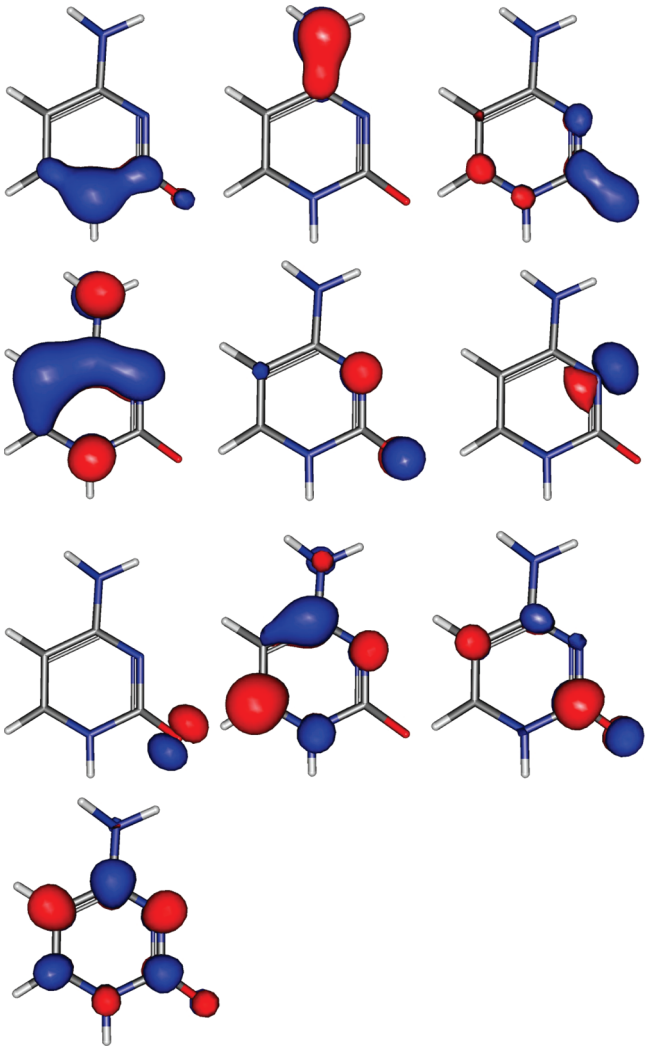


***Figure 3.*** Active orbitals of cytosine.

***Table 5.*** Vertical Excitation Energies $T_v$ (eV) for the $\pi \rightarrow \pi^*$ Transition in Cytosine

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI | exptl $T_0$[91,92] |
|---|---|---|---|---|---|
| 6-31G* | 5.02 | 4.86 | 4.92 | 5.08 | 4.60 |
| 6-31+G* | 4.93 | 4.77 | 4.85 | 5.02 | |
| 6-311G* | 5.03 | 4.87 | 4.92 | 5.10 | |
| 6-311+G* | 4.92 | 4.70 | 4.86 | 5.02 | |

primary chromophores that absorb UV radiation, leading to DNA damage. The ground state molecular structure of cytosine (Figure 2) is almost planar, with the exception of the pyramidal nitrogen atom N1. For the preliminary state-averaged CASSCF calculations, the choice of active space comprised 14 electrons distributed among 10 orbitals (10 $\pi$ orbitals and two lone pair orbitals situated on N8 and O7 in Figure 3). State averaging of the CASSCF calculation was performed over the ground state and the three valence excited states with equal weights. The calculated first excited state at the CASSCF level is a $\pi \rightarrow \pi^*$ state followed by $n_O \rightarrow \pi^*$ and $n_N \rightarrow \pi^*$ excited states involving the lone pair orbitals on the heteroatoms. Tables 5−7 summarize the calculated vertical excitation energies relative to the electronic ground state, in which we see that the energetic ordering of the excited states is already reproduced at the CASSCF level

***Table 6.*** Vertical Excitation Energies $T_v$ (eV) for the $n_O \rightarrow \pi^*$ Transition in Cytosine[a]

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI |
|---|---|---|---|---|
| 6-31G* | 5.36 | 5.13 | 5.05 | 5.28 |
| 6-31+G* | 5.21 | 4.95 | 4.99 | 5.18 |
| 6-311G* | 5.34 | 5.11 | 5.08 | 5.30 |
| 6-311+G* | 5.20 | 4.94 | 5.01 | 5.18 |

[a] The experimental vertical excitation energy for this transition is not available.

***Table 7.*** Vertical Excitation Energies $T_v$ (eV) for the $n_N \rightarrow \pi^*$ Transition in Cytosine

| basis | CASSCF | CASPT2 | MRSDCI | LMRSDCI | exptl $T_0$[91,92] |
|---|---|---|---|---|---|
| 6-31G* | 5.75 | 5.65 | 5.56 | 5.90 | 5.20 |
| 6-31+G* | 5.70 | 5.50 | 5.45 | 5.85 | |
| 6-311G* | 5.75 | 5.66 | 5.55 | 5.93 | |
| 6-311+G* | 5.71 | 5.49 | 5.48 | 5.82 | |

for the two excitation energies that have been measured. The vertical excitation energies calculated at the MRSDCI and LMRSDCI levels for $\pi \rightarrow \pi^*$ and $n_O \rightarrow \pi^*$ typically agree to within 0.2 eV, but the agreement degrades to about 0.4 eV for the $n_N \rightarrow \pi^*$ state. A balanced treatment of differential electron correlation effects is necessary for the evaluation of reliable vertical excitation energies, and a larger correlation domain may be necessary in this instance. Compared to the 2p lone pair on O7, the lone pair on N8 has more 2s character and therefore may be more localized by virtue of being more tightly bound. There is then a greater change for the electron in N8 when going from the $sp^2$ lone pair to a more delocalized $\pi^*$ orbital.

Using a larger correlation domain for both the 2sp and $\pi^*$ orbitals could give a more balanced treatment in this instance, and we explore the effect of varying the local truncation parameters (radius multipliers) on the calculated vertical excitation energies using the 6-31+G* basis. The results are summarized in Table 8, which also contains the computational time, average domain size, and the total number of CSFs for each calculation. While the calculated vertical excitation energies are largely in line with values reported in the earlier tables, a significant improvement is seen for the $n_N \rightarrow \pi^*$ state when less restrictive cutoffs are used for the active orbitals (second column in Table 8) and the discrepancy between MRSDCI and LMRSDCI vertical excitation energies narrows to 0.17 eV. Further inclusion of more correlated electron pairs along with the expansion of the correlation domain (third and fourth column in Table 8) leads to even closer agreement between the MRSDCI and LMRSDCI values, but this is at the expense of higher computational cost.

Finally, we apply our LMRSDCI approach to study the $n_O \rightarrow \pi^*$ transition in $C_{18}H_{36}O$ (Figure 4), which is too costly for a nonlocal MRSDCI treatment. The active space comprises four electrons (nonbonding lone pair on oxygen + two $\pi$ electrons) distributed among three orbitals (p orbital on oxygen + $\pi$ and $\pi^*$ orbitals). The vertical excitation energies are listed in Table 9. The localized nature of the $n_O \rightarrow \pi^*$ excitation within $C_{18}H_{36}O$ is similar to that of acetone, and the calculated vertical excitation energies in these two cases are comparable (cf. Table 4).

***Table 8.*** Vertical Excitation Energies $T_v$ (eV) in Cytosine Using a 6-31+G* Basis As a Function of Local Truncation Parameters[a]

| excitation | MRSDCI | LMRSDCI[b] | LMRSDCI[c] | LMRSDCI[d] |
|---|---|---|---|---|
| $\pi\rightarrow\pi^*$ | 4.85 (288, 122, 5693419) | 5.02 (0.12, 0.45, 0.04) | 4.97 (0.42, 0.56, 0.20) | 4.93 (0.52, 0.74, 0.46) |
| $n_O\rightarrow\pi^*$ | 4.99 (321, 122, 6631128) | 5.16 (0.18, 0.45, 0.07) | 5.12 (0.44, 0.56, 0.22) | 5.11 (0.55, 0.74, 0.48) |
| $n_N\rightarrow\pi^*$ | 5.45 (305, 122, 6174262) | 5.62 (0.18, 0.45, 0.07) | 5.58 (0.46, 0.56, 0.24) | 5.56 (0.59, 0.74, 0.51) |

[a] The CPU time (hours), average domain size, and number of CSFs, respectively, are given in parentheses within the first column. For subsequent columns, the values in parentheses are expressed as a ratio of the corresponding values in the first column. Radius multipliers used for spheres associated with different orbitals are listed in the following order: Inactive, active, virtual, and orbitals residing on heteroatoms. [b] 1.5, 2.0, 1.2, and 3.0. [c] 2.0, 2.0, 1.2 and 3.0. [d] 2.0, 2.0, 2.0 and 3.0.

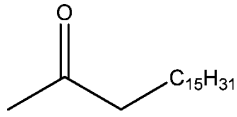



***Figure 4.*** $C_{18}H_{36}O$, a test molecule to study the $n_O\rightarrow\pi^*$ transition using our LMRSDCI approach.

***Table 9.*** Vertical Excitation Energies $T_v$ (eV) for the $n_O\rightarrow\pi^*$ Transition in $C_{18}H_{36}O$

| basis | CASSCF | LMRSDCI |
|---|---|---|
| 6-31G* | 4.51 | 4.62 |
| 6-31+G* | 4.46 | 4.68 |
| 6-311G* | 4.49 | 4.61 |
| 6-311+G* | 4.47 | 4.70 |

## 5. Conclusions

We have investigated here whether local configuration interaction techniques can be used not only for ground states but also for excited states. To this end, we have extended our LMRSDCI method to solve for excited states and then evaluated valence excited states in organic molecules using LMRSDCI. To isolate the interior eigenvalues associated with the excited states of interest, a two-pronged approach was followed. First, the interior eigenvalues are transformed and shifted to the extrema of the spectrum using oblique projection along with a matrix shift within Davidson diagonalization. In this way, we obtain a refined guess for the wave function without solving for the lower roots. Subsequently, residual minimization via the RMM-DIIS method is used for final convergence toward the eigenstate of interest. In most cases, calculated vertical excitation energies within LMRSDCI are within 0.2 eV or better of MRSDCI values, suggesting that good estimates for excited state energetics in large molecules can be obtained with this method, provided that the excitation is fairly localized in nature. Working under the premise of local correlation, we expect that the LMRSDCI approach will be less reliable for excited states that are inherently diffuse (e.g., Rydberg states) and long-ranged in nature (e.g., charge transfer states). However, charge transfer processes that are spatially less extended should be treatable, since the local truncation parameters such as the radius multipliers used in our work may be tuned so that they are large enough to include excitations to/from nearest neighbors. This would allow us to study the shorter-ranged charge transfer processes that would remain a challenge for TDDFT.

We must note in closing that if a large active space is required to model the physical problem, the preceding CASSCF calculation can become the overall bottleneck for the entire calculation, which will limit the applicability of our method. Although the EOM-CC and LR-CC theories do not suffer from this limitation, current low order implementations of these latter methods remain most suitable for describing single electron excitations from ground to excited states that are separately well-described using a single configuration. Multireference methods such as the one presented here are still necessary for all phenomena and molecules in which near-degeneracies appear.

**Acknowledgment.** We are grateful to the National Science Foundation for support of this work. One of the authors (T.S.C) thanks the Agency for Science, Technology and Research A*STAR for funding. We also thank Dr. Johannes Hachmann for pointing out his very helpful related work to us at the beginning of this project.

### References

(1) Kraemer, K. H. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11.

(2) Mukhtar, H.; Elmets, C. A. *Photochem. Photobiol.* **1996**, *63*, 355.

(3) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21.

(4) Grafia, A. M.; Lee, T. J.; Head-Gordon, M. *J. Phys. Chem.* **1995**, *99*, 3493.

(5) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *350*, 522.

(6) Maurice, D.; Head-Gordon, M. *J. Phys. Chem.* **1996**, *100*, 6131.

(7) Casanova, D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *129*, 064104.

(8) Runge, E.; Gross, E. K. U. *Phys. Rev.* **1984**, *136*, B864.

(9) Marques, M. A. L. Time-Dependent Density Functional Theory. In *Lecture Notes in Physics*; Ullrich, C., Nogueira, F., Rubio, A., Gross, E. K. U., Eds.; Springer: Berlin, 2006; Vol. 706, pp 323.

(10) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O. *Mol. Phys.* **1999**, *97*, 859.

(11) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(12) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.

(13) Casida, M. E. *THEOCHEM* **2009**, *914*, 3.

(14) Wu, Q.; Van Voorhis, T. *Phys. Rev. A* **2005**, *72*, 024502.

(15) Wu, Q.; Van Voorhis, T. *J. Chem. Theory Comput.* **2006**, *2*, 765.

(16) Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *125*, 164105.

(17) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57.

(18) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029.

(19) Monkhorst, H. J. *Int. J. Quantum Chem. Symp.* **1977**, *11*, 421.

(20) Dalgaard, E.; Monkhorst, H. J. *Phys. Rev. A* **1983**, *28*, 1217.

(21) Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333.

(22) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 7041.

(23) Scuseria, G. E.; Schaefer, H. F., III. *Chem. Phys. Lett.* **1988**, *152*, 382.

(24) Oliphant, N.; Adamowicz, L. *J. Chem. Phys.* **1992**, *95*, 6645.

(25) Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **1992**, *97*, 4282.

(26) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.

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

(28) Koch, H.; Christiansen, O.; Jørgensen, P.; Sanchez de Merás, A. M.; Helgaker, T. J. *Chem. Phys.* **1997**, *106*, 1808.

(29) Hirata, S. *J. Chem. Phys.* **2004**, *121*, 244106.

(30) Christiansen, O.; Koch, H.; Jørgensen, P.; Olsen, P. J. *Chem. Phys. Lett.* **1996**, *256*, 185.

(31) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *338*, 375.

(32) Krylov, A. I. *Acc. Chem. Res.* **2006**, *39*, 83.

(33) Roos, B. O.; Taylor, P. R. *Chem. Phys.* **1980**, *48*, 157.

(34) Roos, B. O. *Adv. Chem. Phys.* **1987**, *69*, 399.

(35) Malmqvist, P. A.; Rendell, A.; Roos, B. O. *J. Phys. Chem.* **1990**, *94*, 5477.

(36) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483.

(37) Anderson, K.; Malmqvist, P. A.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.

(38) Malmqvist, P. A.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.

(39) Serrano-Andrés, L.; Merchán, M.; Nebot-Gil, I.; Lindh, R.; Roos, B. O. *J. Chem. Phys.* **1993**, *98*, 3151.

(40) Roos, B. O.; Serano-Andrés, L.; Merchán, M. *Pure Appl. Chem.* **1993**, *65*, 1693.

(41) Roos, B. O.; Fülscher, M. P.; Malmqvist, P. A.; Merchán, M.; Serrano-Andrés, L. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 357.

(42) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083.

(43) Malmqvist, P. A.; Pierloot, K.; Rehaman Moughal Shahi, A.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.

(44) Partridge, H.; Langhoff, S. R.; Bauschlicher, C. W., Jr. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 209.

(45) Korona, T.; Werner, H.-J. *J. Chem. Phys.* **2003**, *118*, 3006.

(46) Crawford, T. D.; King, R. A. *Chem. Phys. Lett.* **2002**, *366*, 611.

(47) Kats, D.; Korona, T.; Schütz, M. *J. Chem. Phys.* **2006**, *125*, 104106.

(48) Kats, D.; Korona, T.; Schütz, M. *J. Chem. Phys.* **2007**, *127*, 064107.

(49) Kats, D.; Schütz, M. *J. Chem. Phys.* **2009**, *131*, 124117.

(50) Kats, D.; Schütz, M. *Z. Phys. Chem.* **2010**, *224*, 601.

(51) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151.

(52) Saebø, S.; Pulay, P. *Chem. Phys. Lett.* **1985**, *113*, 13.

(53) Pulay, P.; Saebø, S. *Theor. Chim. Acta* **1986**, *69*, 357.

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

(55) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1988**, *88*, 1884.

(56) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.

(57) Paige, C. C.; Parlett, B. N.; van der Vorst, H. A. *Numer. Linear Algebra Appl.* **1995**, *2*, 115.

(58) Morgan, R. B. *Linear Algebra Appl.* **1991**, *154*, 289 (1991).

(59) Sleijpen, G. L. G.; van der Vorst, H. A. *SIAM J. Matrix Anal. Appl.* **1996**, *17*, 401.

(60) Dorando, J. J.; Hachmann, J.; Chan, G. K.-L. *J. Chem. Phys.* **2007**, *127*, 084109.

(61) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393.

(62) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.

(63) Csaszar, P.; Pulay, P. *J. Mol. Struct.* **1984**, *114*, 31.

(64) Wood, D. M.; Zunger, A. *J. Phys. A: Math. Gen.* **1985**, *18*, 1343.

(65) Hutter, J.; Lüthi, H. P.; Parrinello, M. *Comput. Mater. Sci.* **1994**, *2*, 244.

(66) Kresse, G.; Furthmuller, J. *Phys. Rev. B* **1996**, *54*, 11169.

(67) Whitten, J. L.; Hackmeyer, M. *J. Chem. Phys.* **1969**, *51*, 5584.

(68) Peyerimhoff, D.; Buenker, J. *Chem. Phys. Lett.* **1972**, *16*, 235.

(69) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87.

(70) Liu, B. *Numerical Algorithm in Chemistry: Algebraic Methods*; LBL-8158, UC-32, CONF-780878; National Resource for Computation in Chemistry, Lawrence Berkeley Laboratory: Berkeley, California, 1978; pp 49.

(71) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

(72) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P. A.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2009**, *31*, 224.

(73) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916.

(74) Subotnik, J. E.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2005**, *123*, 114108.

(75) Bytautas, L.; Ivanic, J.; Ruedenberg, K. *J. Chem. Phys.* **2003**, *119*, 8217.

(76) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683.

(77) Røeggen, I.; Wisløff-Nielsen, E. *Chem. Phys. Lett.* **1986**, *132*, 154.

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

(79) Reynolds, G.; Martinez, T. J.; Carter, E. A. *J. Chem. Phys.* **1996**, *105*, 6455.

(80) Reynolds, G.; Carter, E. A. *Chem. Phys. Lett.* **1997**, *265*, 660.

(81) Walter, D.; Carter, E. A. *Chem. Phys. Lett.* **2001**, *346*, 177.

(82) Walter, D.; Szilva, A. B.; Niedfeldt, K.; Carter, E. A. *J. Chem. Phys.* **2002**, *117*, 1982.

(83) Walter, D.; Venkatnathan, A.; Carter, E. A. *J. Chem. Phys.* **2003**, *118*, 8127.

(84) Chwee, T. S.; Szilva, A. B.; Lindh, R.; Carter, E. A. *J. Chem. Phys.* **2008**, *128*, 224106.

(85) Chwee, T. S.; Carter, E. A. *J. Chem. Phys.* **2010**, *132*, 074104.

(86) Malmqvist, P. A.; Roos, B. O.; Fülscher, M. P.; Rendell, A. *Chem. Phys.* **1992**, *162*, 359.

(87) Roos, B. O.; Andersson, K.; Fülscher, M. P. *Chem. Phys. Lett.* **1992**, *192*, 5.

(88) Angeli, C. *J. Comput. Chem.* **2009**, *30*, 1319.

(89) Hiraya, A.; Shobatake, K. *J. Chem. Phys.* **1991**, *94*, 7700.

(90) Mulliken, R. S. *J. Chem. Phys.* **1935**, *8*, 564.

(91) Clark, L. B.; Peschel, G. G.; Tinoco, I. *J. Phys. Chem.* **1965**, *69*, 3615.

(92) Clark, L. B.; Tinoco, I. *J. Am. Chem. Soc.* **1965**, *87*, 11.

CT100486Q

JCTC Journal of Chemical Theory and Computation

# G4(MP2)-6X: A Cost-Effective Improvement to G4(MP2)

Bun Chan,* Jia Deng,† and Leo Radom*

*School of Chemistry and ARC Center of Excellence for Free Radical Chemistry and Biotechnology, University of Sydney, Sydney, NSW 2006, Australia*

Received September 21, 2010

**Abstract:** G4(MP2)-6X is developed as a composite procedure with a cost comparable to that of G4(MP2) but performance approaching that of G4. The new procedure is a variant of G4(MP2) that employs BMK/6-31+G(2df,p) geometries and has six additional scaling factors for the correlation energy components. The scaling factors and HLC parameters are optimized using the new E2 set of 526 energies, representing thermochemical properties, reaction energies and barriers, and weak interactions. G4(MP2)-6X achieves a mean absolute deviation (MAD) from benchmark values of 3.64 kJ mol$^{-1}$ for the E2 set, compared with 4.42 kJ mol$^{-1}$ for G4(MP2). For the E0 set of 148 energies, G4(MP2)-6X gives an MAD of 3.43 kJ mol$^{-1}$, compared with 3.22 kJ mol$^{-1}$ for G4 and 4.03 kJ mol$^{-1}$ for G4(MP2). The new G4(MP2)-6X procedure thus uses extra parametrization to provide a G4-type performance without incurring G4-type computational costs.

## 1. Introduction

Composite quantum chemistry methods[1] enable the accurate prediction of thermochemical properties at reduced cost in terms of computing resources. A variety of such procedures have been developed, including the Gaussian (Gn) procedures,[2] the complete-basis-set (CBS) methods,[3] and the Weizmann (Wn) procedures.[4] These methods range from highly accurate at modest computational cost to exceptionally accurate at much higher expense. For instance, G4 achieves a mean absolute deviation (MAD) from experimental values of 3.47 kJ mol$^{-1}$ (0.83 kcal mol$^{-1}$) for the G3/05 set[5] of 454 energies, using MP2, MP4, and CCSD(T) component calculations of moderate individual cost. At the other end of the spectrum, by employing coupled cluster calculations up to CCSDTQ5, W4[4c] gives an MAD of just 0.42 kJ mol$^{-1}$ (0.1 kcal mol$^{-1}$) for the W3 set[4b] of 36 energies.

For many chemical systems of medium size, the Wn procedures, and to some extent the Gn procedures, are still prohibitively expensive. To address this problem, the Gn-(MP2) procedures have been formulated.[6] These variants of Gn are less computationally demanding and are therefore applicable to a much wider range of systems. The trade-off for the reduced computational requirements is that the Gn(MP2) procedures are somewhat less accurate than the Gn methods. For example, G4(MP2)[6c] has an MAD of 4.35 kJ mol$^{-1}$ (1.04 kcal mol$^{-1}$) for the G3/05 set, compared with 3.47 kJ mol$^{-1}$ for G4.

The broader-ranging applicability of G4(MP2) makes it attractive to seek further improvement in its performance. Is it feasible to achieve G4-type performance while retaining G4(MP2)-type computational cost, perhaps through additional parametrization? In the present study, we examine various modifications of G4(MP2) with this end in mind and propose G4(MP2)-6X as a cost-effective improved procedure.

## 2. Computational Details

Standard ab initio molecular orbital theory and density functional theory (DFT) calculations[7] were carried out with the Gaussian 03,[8] Gaussian 09,[9] Molpro 2006,[10] and Psi 3[11] programs. Unless otherwise noted, geometries were optimized with the BMK[12] DFT procedure using the 6-31+G(2df,p) basis set. In specific cases, zero-point vibrational energies (ZPVEs) and thermal corrections to enthalpy ($\Delta H$) at 298 K, derived from scaled BMK/6-31+G(2df,p) frequencies, were incorporated into the total energies.[13,14] Single-point energies were obtained at the HF, MP2, and CCSD(T) levels for the composite procedures. Unless otherwise noted,

* Corresponding authors e-mail: chan_b@chem.usyd.edu.au; radom@chem.usyd.edu.au.

† Current address: Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia.

10.1021/ct100542x 
Published on Web 12/08/2010

energies are reported in kilojoules per mole. In order to maximize both the accuracy and precision of the new procedures, i.e., to minimize the deviations (maximize accuracy) and the variations in the deviations (maximize precision), we simultaneously minimize the mean absolute deviation (MAD) from the benchmark values, and the standard deviation (SD) of the deviations. Preliminary calculations indicate that this is advantageous compared with the typical approach of minimizing the MAD alone. Thus, the MAD/SD procedure leads to a lower SD and yields fewer outliers (deviations > 8.37 kJ $mol^{-1}$ (2 kcal $mol^{-1}$)) without changing the MAD by more than 0.01 kJ $mol^{-1}$. In the course of these investigations, we have briefly examined various approaches for the simultaneous minimization of the MAD and SD, including minimizing the mean of the MAD and SD values [1/2 × (MAD + SD)], their root-mean-square value [$2^{-1/2}$ × $(MAD^2 + SD^2)^{1/2}$], and their harmonic mean [2 × $(MAD^{-1} + SD^{-1})^{-1}$]. We find that the three averaging procedures give very similar results, and we have chosen for simplicity to use the mean of MAD and SD in our parametrization processes.

## 3. Theory

**3.1. Description of G4(MP2).** The G4(MP2) procedure seeks to approximate a high-level energy (CCSD(T) with a large basis set) through the use of a series of lower-level energies and an additivity scheme. While the details of the theory have been described in the original paper,[6c] we provide here a brief summary for comparison purposes. The G4(MP2) energy is obtained in the following manner:

$$\text{G4(MP2)} = \text{HF/CBS} + E_{\text{MP2}}^{\text{corr}}/\text{G3MP2LargeXP} + \Delta E_{\text{CCSD(T)}}/\text{6-31G(d)} + \text{HLC} + \text{ZPVE} + E_{\text{SO}} \quad (1)$$

• The geometry is obtained at the B3-LYP/6-31G(2df,p) level.

• HF/CBS is an estimate of the Hartree−Fock-limit energy and is obtained by extrapolation to the complete-basis-set limit with aug-cc-pV(n+d)Z (n = T, Q) basis sets modified by reducing the number of diffuse and polarization functions, using the formula $E_{\text{CBS}} = [E_{\text{Q}} - E_{\text{T}} \exp(-1.63)]/[1 - \exp(-1.63)]$.

• The frozen-core approximation is employed for all correlation calculations. More specifically, the largest noble-gas core is frozen, except that the following are treated as valence orbitals:

◦ 3d orbitals on third-row main-group elements (Ga−Kr)

◦ 2s and 2p orbitals on Na and Mg and 3s and 3p orbitals on K and Ca

• $E_{\text{MP2}}^{\text{corr}}$ is the MP2 correlation energy (i.e., $E_{\text{MP2}} - E_{\text{HF}}$), calculated with the G3MP2LargeXP basis set.

• $\Delta E_{\text{CCSD(T)}}$ is the CCSD(T) correlation energy beyond MP2 (i.e., $E_{\text{CCSD(T)}}^{\text{corr}} - E_{\text{MP2}}^{\text{corr}}$), calculated with the 6-31G(d) basis set.

• HLC is a higher-level-correction term that depends on the number of α and $\beta$ valence electrons ($n_\alpha$ and $n_\beta$). It is obtained with parameters ($A$, $A'$, $B$, $C$, $D$, and $E$, mHartree) optimized using the G3/05 set:

◦ $-A\ n_\beta$ for closed-shell molecules

◦ $-A'\ n_\beta - B\ (n_\alpha - n_\beta)$ for open-shell molecules

◦ $-C\ n_\beta - D\ (n_\alpha - n_\beta)$ for atomic species

◦ $-E\ n_\beta$ for "single electron pair" species (e.g., $Li_2$)

◦ $A = 9.472$, $A' = 9.769$, $B = 3.179$, $C = 9.741$, $D = 2.115$, $E = 2.379$

• The zero-point vibrational energy (ZPVE) and thermal corrections to enthalpy are obtained with scaled (0.9854) B3-LYP/6-31G(2df,p) frequencies.

• A spin−orbit correction ($E_{\text{SO}}$), where available from experimental results or from accurate calculations, is included.

The sum of the first three terms in eq 1, HF/CBS + $E_{\text{MP2}}^{\text{corr}}$/G3MP2LargeXP + $\Delta E_{\text{CCSD(T)}}$/6-31G(d), is designed to approximate the electronic energy at the CCSD(T)/CBS level. The HLC, through parametrization to experimental data, is used to compensate for the remaining deficiencies in the theory and in the additivity scheme.

The form of G4(MP2) indicates several components that determine its accuracy. Needless to say, an important factor is the quality of the electronic structure methods that are used for geometry optimization and in the approximation of CCSD(T)/CBS. The HLC has no effect on the energy of reactions that involve closed-shell reactants and closed-shell products. However, for radical reactions and some properties such as heats of formation, ionization energies, and electron affinities, the HLC can substantially improve the result. Thus, both the form of the HLC and the experimental data employed in the fitting are important for many properties and reactions.

**3.2. G4(MP2)-6X: An Alternative Composite Procedure.** While the use of the HLC in G4(MP2) is generally very successful in improving the prediction of thermochemical properties, other approaches have also been employed to improve the performance of a simple additivity scheme. For instance, the G3S[15] and MCCM/3[16] procedures scale the energy of the various components in the additivity scheme, in place of using HLC corrections. In the present study, we introduce the G4(MP2)-6X procedure, which includes such scaling as well as other modifications designed to provide an improvement to G4(MP2) but without significantly affecting the computational cost.

G4(MP2)-6X has six extra parameters that are independent of the number of electrons, in addition to the six HLC parameters in the original G4(MP2). These are introduced in the hope that they will not only improve the general accuracy of the procedure but also provide an improvement in the cases where the HLC does not contribute to the relative energies as, for example, for barriers for closed-shell reactions.

The form of the six additional parameters is loosely based on the G3S procedure,[15] where the correlation contributions for MP2, CCSD, and the perturbative triples correction (T) are separately scaled. In addition, we make use of a modified MP2 method, namely, SCS-MP2, which has been found by Grimme to show improved performance over standard MP2.[17] This is achieved by splitting the MP2 correlation energy ($E_{\text{MP2}}^{\text{corr}}$) into opposite-spin (OS) and same-spin (SS) contributions and applying individual scaling to the two components.[17] In G4(MP2)-6X, we also scale the OS and

SS contributions to $E_{MP2}^{corr}$ separately. With a similar philosophy to that for SCS-MP2, Sherrill and co-workers have introduced the SCS-CCSD method.[18] However, our preliminary investigations show that separate scaling of the OS and SS contributions to the CCSD correlation energy, in the context of G4(MP2)-type procedures, does not lead to a noticeable improvement, and we therefore do not apply SCS-CCSD to G4(MP2)-6X. Finally, we use an improved procedure for geometry optimization (BMK/6-31+G(2df,p)). The description of the G4(MP2)-6X procedure is then as follows:

$$\begin{aligned} \text{G4(MP2)-6X} = \; & \text{HF/CBS} + E_{\text{SCS-MP2}}^{\text{corr}}/\text{G3MP2LargeXP} + \\ & \Delta E_{\text{S-CCSD}}/\text{6-31G(d)} + E_{\text{S-(T)}}^{\text{corr}}/\text{6-31G(d)} + \\ & \text{HLC} + \text{ZPVE} + E_{\text{SO}} \end{aligned} \tag{2}$$

• The geometry is obtained at the BMK/6-31+G(2df,p) level.

• HF/CBS is obtained in the same manner as for G4(MP2).[19]

• The following correlation energies are obtained with the G4 frozen core:

◦ $c1 \cdot E_{MP2OS}^{corr}$ and $c2 \cdot E_{MP2SS}^{corr}$ are the scaled OS and scaled SS contributions to the MP2/6-31G(d) correlation energy, respectively.

◦ $c3 \cdot E_{MP2OS}^{corr\prime}$ and $c4 \cdot E_{MP2SS}^{corr\prime}$ are the scaled OS and scaled SS contributions to the MP2/G3MP2LargeXP correlation energy, respectively.

◦ $c5 \cdot E_{CCSD}^{corr}$ is the scaled CCSD contribution to the CCSD(T)/6-31G(d) correlation energy.

◦ $c6 \cdot E_{(T)}^{corr}$ is the scaled perturbative triples contribution to the CCSD(T)/6-31G(d) correlation energy.

• $E_{SCS-MP2}^{corr}/\text{G3MP2LargeXP} = c3 \cdot E_{MP2OS}^{corr\prime} + c4 \cdot E_{MP2SS}^{corr\prime}$

• $\Delta E_{S-CCSD}/\text{6-31G(d)} = c5 \cdot E_{CCSD}^{corr} - (c1 \cdot E_{MP2OS}^{corr} + c2 \cdot E_{MP2SS}^{corr})$

• $E_{S-(T)}^{corr}/\text{6-31G(d)} = c6 \cdot E_{(T)}^{corr}$

• c1, c2, c3, c4, c5, and c6 are parameters optimized using the E2 training set; see section 3.3 for details of E2.

• The HLC has the same form as that for G4(MP2), but the HLC parameters are reoptimized.

• Scaled BMK/6-31+G(2df,p) frequencies are used to obtain the ZPVE (0.9770) and thermal corrections to enthalpy (0.9627).[13,14]

• The same spin−orbit contribution as in G4(MP2) is included.

A script for running G4(MP2)-6X calculations using Gaussian 09 is included in the Supporting Information.

We note that several singlet species ($C_2$, $Be_2$, and $CN^+$) among our training sets exhibit RHF to UHF instabilities. This suggests that they have substantial biradical character and are not adequately described by the RHF wave functions. We find that in these cases, the use of an unrestricted wave function leads to improved agreement with benchmark thermochemistry. However, the broken-spin-symmetry UHF wave functions in these cases are found to be highly spin-contaminated. While the improvements with UHF are encouraging, a multireference treatment is likely to help further. Nonetheless, it is advisible that an initial stability test on the reference wave function be conducted to identify potential problems of this type, and to provide some improvement.

**3.3. Training and Test Sets.** In developing new composite procedures, training sets are employed to optimize the empirical parameters, while test sets are used to evaluate their performance. There are numerous training and test sets that have been previously used. For example, the G3/05 set[5] employed for the parametrization of the HLC in G4(MP2) comprises 270 heats of formation, 105 ionization energies, 63 electron affinities, 10 proton affinities, and six hydrogen-bond energies. On the other hand, the performance of W4 was assessed with the W3 set[4b] of 36 atomization energies.

While the above and many other training and test sets consist of data for general thermochemistry, there have also been a number of specialized sets developed for specific energy properties, such as barriers and weak interactions. Thus, the DBH24 set[20] consists of 24 barriers, obtained at the W1−W4 levels of theory, for a diverse range of reaction types. These include hydrogen abstractions, $S_N2$ reactions, hydrogen transfers, and unimolecular decompositions. Using W2 calculations, Boese et al. have compiled a hydrogen-bonding set of 16 complexes[21] (referred to as the HB16 set hereafter), while the WI9/04 set[22] comprises complexation energies for nine weakly bound complexes, calculated at the W1 level. Very recently, the GMTKN24 data set has been proposed,[23] which comprises subsets of a diverse range of 24 chemical test sets, encompassing thermochemistry, kinetics, and noncovalent interactions. The GMTKN24 set contains 731 energies.

In this study, we propose two new training and test sets that include a variety of thermochemical properties. They are termed the "energy sets" E0 and E2 and contain 148 and 526 energies, respectively. These sets contain experimental and theoretical energies of chemical accuracy. We use the E0 set for evaluating methods for geometry optimization and the E2 set for training the G4(MP2)-6X procedure. We also examine the use of the "traditional" and moderately sized G2/97 set (302 energies) for parametrization.[24] The details of the training and test sets are given below.

The E0 set is a selected subset of highly accurate theoretical data presented recently by Karton et al.[25] It contains atomization energies (AE, W4/08), barriers ($\Delta H^{\ddagger}$, DBH24), hydrogen-bond energies (HB, HB16), and energies for weakly bound complexes (BE, WI9/04):

$$\begin{aligned} \text{E0(148)} = \; & \text{W4/08 (99, 0 K, W4+)} \\ & + \text{DBH24 (24, vibrationless, W3+)} \\ & + \text{HB16 (16, vibrationless, W2)} \\ & + \text{WI9/04 (9, vibrationless, W1)} \end{aligned}$$

Shown in parentheses are the number and the specification of the energies, and the methods by which they are obtained. W3+ and W4+ signify theoretical values obtained with procedures of at least W3 and W4 quality, respectively.

While the details of the G2/97 set have been given elsewhere,[24] we provide here a brief summary for comparison purposes. It consists of experimental data for heats of formation ($\Delta H_f$), ionization energies (IEs), electron affinities (EAs), and proton affinities (PAs):

$$\begin{aligned}\text{G2/97(302)} = \ & \text{G2/97}\Delta H_f\,(148, 298\ \text{K, expt}) \\ & + \text{G2/97 IE (88, 0 K, expt)} \\ & + \text{G2/97 EA (58, 0 K, expt)} \\ & + \text{G2/97 PA (8, 0 K, expt)}\end{aligned}$$

For the IE of CN in this set, an alternative experimental value[26] (14.03 ± 0.02 eV, 1353 kJ mol$^{-1}$) has been recommended on the basis of W1 and W2 calculations.[27] We have adopted this value in the present study.

Our largest set is the E2 set, which is comprised of the E0 set, a modified G2/97 set termed G2/97′, and the new E1 set:

$$\text{E2 (526)} = \text{E0 (148)} + \text{G2/97}' \text{ (248)} + \text{E1 (130)}$$

The G2/97′ set reduces the number of heats of formation from 148 (in G2/97) to 94. The new heats of formation set in G2/97′ is simply termed G2/97′ $\Delta H_f$. The reduction in the number of data is done to remove redundancies in the final E2 set, as 54 of the 148 molecules in G2/97 are also present in the W4/08 set. Thus, the G2/97′ set is defined as

$$\begin{aligned}\text{G2/97}'\text{(248)} = \ & \text{G2/97}'\Delta H_f\,(94, 298\ \text{K, expt}) \\ & + \text{G2/97 IE (88, 0 K, expt)} \\ & + \text{G2/97 EA (58, 0 K, expt)} \\ & + \text{G2/97 PA (8, 0 K, expt)}\end{aligned}$$

The E1 set consists of 73 heats of formation from the G3/99 set[28] that are not present in G2/97, calculated enthalpies for 49 radical reactions of Coote et al.[29] (referred to hereafter as RR49), and a subset (referred to as PR8) of the pericyclic reaction set of Houk et al.[30]

G3/99 introduces 75 additional $\Delta H_f$ values to G2/97, of which two are already present in the W4/08 set. Removal of this duplication leads to the G3/99′ $\Delta H_f$ set. The RR49 set comprises enthalpies calculated at the W1 level for 21 radical-addition (ADD) and 28 hydrogen-abstraction (ABS) reactions. Houk et al.'s pericyclic reaction set consists of experimental values for 11 barriers and 7 enthalpies for 11 pericyclic reactions. However, due to uncertainties in the experimental values, we choose to employ a subset of eight barriers, for which W1 vibrationless values are available.[25] The E1 set is thus defined as

$$\begin{aligned}\text{E1(130)} = \ & \text{G3/99}'\Delta H_f\,(73, 298\ \text{K, expt}) \\ & + \text{ADD (21, 0 K, W1, part of RR49)} \\ & + \text{ABS (28, 0 K, W1, part of RR49)} \\ & + \text{PR8 (8, vibrationless, W1)}\end{aligned}$$

## 4. Results and Discussion

We have optimized the parameters for G4(MP2)-6X using our largest E2 set (Table 1). We note that most of the HLC parameters for G4(MP2)-6X are smaller than their G4(MP2) counterparts. This is advantageous because it means that the new method is less reliant on the HLC than the original implementation. It has been pointed out that the form of HLC in G4(MP2), specifically the use of different values for $A$ and $A'$ for closed- and open-shell species, can lead to a diverging error for radical reactions.[29,31] However, the difference between $A$ and $A'$ for G4(MP2)-6X (0.091) is substantially smaller than that for G4(MP2) (0.297). As a result, the problem of a diverging HLC contribution to radical reactions is significantly reduced for the new procedure.

***Table 1.*** Optimized HLC ($A-E$, mHartree) and Scaling (c1−c6) Parameters for G4(MP2)-6X[a]

| | G4(MP2)[b] | G4(MP2)-6X | | G4(MP2)-6X |
|---|---|---|---|---|
| $A$ | 9.472 | 7.173 | c1 | 1.327 |
| $A'$ | 9.769 | 7.264 | c2 | 0.403 |
| $B$ | 3.102 | 3.677 | c3 | 1.249 |
| $C$ | 9.741 | 7.239 | c4 | 0.486 |
| $D$ | 2.115 | 2.404 | c5 | 1.077 |
| $E$ | 2.379 | 1.021 | c6 | 0.824 |

[a] Parameterized with the E2 set. See section 3.2 for the definition of the parameters. [b] Reference 6.

In the following sections, we will first examine the performance of various methods for geometry optimization. Next, we will evaluate the new parameters in G4(MP2)-6X and assess their importance in improving the new procedure. We will then look at the performance of G4(MP2)-6X in comparison with G4(MP2), as well as discuss the portability of the popular G2/97 training set by testing the G2/97-trained procedure on the larger and more diverse E2 set. We will conclude our discussion by providing a more detailed analysis of the strengths and shortcomings of G4(MP2)-6X.

**4.1. The Choice of Geometry.** In the formulation of G4, Curtiss et al. have identified cases where closer agreement with benchmark energies can be achieved by using procedures for optimizing the geometry other than the default B3-LYP/6-31G(2df,p) method.[2d] For instance, the use of MP2(Full)/6-31G(d) geometries improves the IEs for $CH_4$, $BF_3$, and $BCl_3$, while hydrogen-bond energies are improved by using B3-LYP/6-31+G(2df,p) geometries.

While the use of MP2 for geometry optimization may be desirable in some cases, it can become too costly to be employed for routine computations for larger systems, and it can also fail for open-shell systems for which there is significant spin contamination. We therefore focus on the use of alternative DFT procedures with various basis sets for obtaining geometries. In particular, we examine the performance of the BMK[12] and M06-2X[32] procedures in combination with the 6-31+G(d,p), 6-31G(2df,p), 6-31+G(2df,p), and 6-311+G(2df,p) basis sets for geometry optimization. We then calculate single-point energies at the G4(MP2) level on these geometries. We use the E0 test set for our assessment, as it contains not only standard molecules, but also transition structures and hydrogen-bonded and weak complexes, which can be expected to be challenging in terms of obtaining reliable geometries. The MAD values for the E0 set and its subsets, for the various procedures used for geometry optimization, are shown in Table 2.

We find that, for the full E0 set and with the largest 6-311+G(2df,p) basis set, both BMK (4.00 kJ mol$^{-1}$) and M06-2X (4.24 kJ mol$^{-1}$) lead to lower MADs than B3-LYP (4.34 kJ mol$^{-1}$). We can see that this is also true for the subsets of E0, with the exception of the W4/08 set, for which the MADs for BMK and M06-2X are 0.29 and 0.63 kJ mol$^{-1}$, respectively, higher than that for B3-LYP. For BMK, we find that the MAD generally becomes smaller as the size of the basis set is increased. On the other hand, the basis set

***Table 2.*** Mean Absolute Deviations (MADs, kJ mol$^{-1}$) from Benchmark Energies As a Function of Methods Employed for Geometry Optimization for G4(MP2)[a]

| | | 6-31+G (d,p) | 6-31G (2df,p) | 6-31+G (2df,p)[b] | 6-311+G (2df,p) |
|---|---|---|---|---|---|
| W4/08 | B3-LYP | | 4.36 | | 4.47 |
| | BMK | 4.98 | 4.90 | 4.72 | 4.76 |
| | M06-2X | 4.95 | 4.99 | 5.09 | 5.10 |
| DBH24 | B3-LYP | | 5.71 | | 5.75 |
| | BMK | 3.51 | 3.05 | 2.91 | 2.72 |
| | M06-2X | 2.65 | 2.93 | 2.69 | 2.60 |
| HB16 | B3-LYP | | 3.17 | | 1.92 |
| | BMK | 1.91 | 2.85 | 1.76 | 1.74 |
| | M06-2X | 2.15 | 3.47 | 2.08 | 2.03 |
| WI9/04 | B3-LYP | | 4.09 | | 3.47 |
| | BMK | 3.88 | 3.26 | 3.35 | 3.20 |
| | M06-2X | 2.89 | 4.32 | 2.90 | 3.08 |
| E0 | B3-LYP | | 4.43 | | 4.34 |
| | BMK | 4.35 | 4.28 | 4.03 | 4.00 |
| | M06-2X | 4.15 | 4.45 | 4.24 | 4.24 |

[a] Literature scale factors were employed for obtaining ZPVEs and $\Delta H$. [b] The appropriate scaling factors for the 6-31+G(2df,p) basis set are taken as the average of the values for 6-31G(2df,p) and 6-311+G(2df,p); see ref 14 for more details.

dependence for M06-2X is more erratic, with the lowest MAD for the E0 set being achieved with the smallest 6-31+G(d,p) basis set, and the largest MAD for the intermediate-sized 6-31G(2df,p) basis set. For the subsets of E0, we can also see a more systematic basis set effect for BMK than for M06-2X.

While both BMK and M06-2X appear to provide superior geometries compared with those of B3-LYP, the somewhat lower MAD for BMK, and the seemingly more predictable basis set effect, have led us to favor BMK as our method of choice. In addition, it has previously been found that the use of BMK geometries improves the agreement with experimental thermochemical data for G3.[33] Turning our attention to the choice of basis set, the use of 6-31+G(2df,p) gives an overall MAD (4.03 kJ mol$^{-1}$) that is only slightly greater than that for the larger 6-311+G(2df,p) basis set (4.00 kJ mol$^{-1}$). However, further reduction in the basis set size to 6-31G(2df,p) or 6-31+G(d,p) leads to larger increases in the MAD (to 4.28 and 4.35 kJ mol$^{-1}$, respectively). Thus, we have chosen BMK/6-31+G(2df,p) as the method for geometry optimization in the G4(MP2)-6X procedure. We note that the use of 6-31+G(2df,p), compared with the smaller 6-31G(2df,p) basis set employed in G4(MP2), will lead to a slightly more expensive procedure. However, the geometry optimization generally contributes only a small part to the total computational cost in a G4(MP2)-type calculation, and this is therefore a cost-effective improvement.

**4.2. Components of G4(MP2)-6X and Their Contributions.** Having determined the appropriate procedure for geometry optimization, we now proceed to examine the various approaches for scaling the correlation energies used in G4(MP2), and how they affect the performance of the modified procedures. The results are summarized in Table 3.

We can see that, when compared with G4(MP2) with the default B3-LYP geometries (column I), the use of BMK for geometry optimization (column II) gives a comparable MAD, a reduced SD, and three less outliers.[34] Reoptimization of the HLC parameters leads to a lower MAD and SD (column III). When one adds a single set of scaling parameters (c1 and c2) for the OS and SS components for the MP2 correlation energy ($E_{\mathrm{MP2}}^{\mathrm{corr}}$), the MAD and SD are further lowered (column IV). However, the biggest improvement accompanying this change can be seen for NO, where the number of outliers is substantially reduced from 68 to 59. A slight improvement in performance can be achieved when one scales c5 and c6 (instead of c1 and c2) for the CCSD and (T) correlation energies, respectively (column VI). Introducing different scaling parameters for the two sets of $E_{\mathrm{MP2}}^{\mathrm{corr}}$, i.e., c1 and c2 for $E_{\mathrm{MP2}}^{\mathrm{corr}}$/6-31G(d) and c3 and c4 for $E_{\mathrm{MP2}}^{\mathrm{corr}}$/G3MP2LargeXP, leads to a larger improvement (column V). Finally, when one applies scaling to all six components, one arrives at the full G4(MP2)-6X procedure (column VII), with the lowest values for the MAD (3.64 kJ mol$^{-1}$) and SD (5.14 kJ mol$^{-1}$), and with a significant reduction also in the number of outliers to 44.

**4.3. Performance of G4(MP2)-6X.** Table 4 summarizes the statistics for the E2 test set for G4(MP2), and for G4(MP2)-6X when parametrized with the G2/97 and E2 training sets. We can see that G4(MP2) gives a respectable MAD of 4.42 kJ mol$^{-1}$, and an SD of 6.38 kJ mol$^{-1}$. The

***Table 3.*** Performance of Various Modifications for G4(MP2) on the E2 Set[a]

| | I | II | III | IV | V | VI | VII |
|---|---|---|---|---|---|---|---|
| Geometry[b] | B3-LYP[c] | BMK | BMK | BMK | BMK | BMK | BMK[d] |
| HLC | default | default | fitted | fitted | fitted | fitted | fitted |
| c1 | 1 | 1 | 1 | fitted | fitted | 1 | fitted |
| c2 | 1 | 1 | 1 | fitted | fitted | 1 | fitted |
| c3 | 1 | 1 | 1 | = c1 | fitted | 1 | fitted |
| c4 | 1 | 1 | 1 | = c2 | fitted | 1 | fitted |
| c5 | 1 | 1 | 1 | 1 | 1 | fitted | fitted |
| c6 | 1 | 1 | 1 | 1 | 1 | fitted | fitted |
| MD[e] | −0.49 | −0.39 | −0.37 | 0.08 | −0.15 | −0.40 | −0.26 |
| MAD[f] | 4.42 | 4.39 | 4.18 | 3.99 | 3.77 | 3.82 | 3.64 |
| LD[g] | −41.61 | −42.17 | −38.14 | −37.79 | −39.41 | −37.45 | −36.66 |
| SD[h] | 6.38 | 6.05 | 5.88 | 5.61 | 5.41 | 5.47 | 5.14 |
| NO[i] | 70 | 67 | 68 | 59 | 55 | 56 | 44 |

[a] Parameterized with the E2 set. See section 3.2 for the definition of the parameters. All energies in kJ mol$^{-1}$. [b] The 6-31G(2df,p) basis set was used for B3-LYP, while BMK employs 6-31+G(2df,p). [c] This set of parameters represents the original G4(MP2). [d] G4(MP2)-6X. [e] Mean deviation. [f] Mean absolute deviation. [g] Largest deviation. [h] Standard deviation. [i] Number of outliers (absolute deviation > 8.37 kJ mol$^{-1}$).

***Table 4.*** Comparison of Deviations (kJ mol$^{-1}$) for the E2 Test Set for G4(MP2), and for G4(MP2)-6X When Parameterized with the G2/97 and E2 Training Sets[a]

| G4(MP2) | | | | | |
|---|---|---|---|---|---|
| test set | MD | MAD | LD | SD | NO |
| E0 | −0.57 | 4.58 | −29.64 | 6.78 | 20 |
| G2/97′ | −0.78 | 4.42 | −41.61 | 6.53 | 37 |
| E1 | 0.16 | 4.22 | −20.15 | 5.55 | 13 |
| E2 | −0.49 | 4.42 | −41.61 | 6.38 | 70 |
| **G4(MP2)-6X (parametrized with G2/97)** | | | | | |
| test set | MD | MAD | LD | SD | NO |
| E0 | −1.34 | 3.81 | −19.40 | 5.05 | 15 |
| G2/97′ | −0.28 | 4.09 | −38.53 | 5.85 | 31 |
| E1 | 0.76 | 3.44 | 15.28 | 4.38 | 9 |
| E2 | −0.32 | 3.85 | −38.53 | 5.34 | 55 |
| **G4(MP2)-6X (parametrized with E2)** | | | | | |
| test set | MD | MAD | LD | SD | NO |
| E0 | −0.42 | 3.43 | −19.47 | 4.67 | 8 |
| G2/97′ | −0.49 | 4.14 | −36.66 | 5.87 | 29 |
| E1 | 0.37 | 2.92 | 15.17 | 4.04 | 7 |
| E2 | −0.26 | 3.64 | −36.66 | 5.14 | 44 |

[a] See section 3.3 for the definitions of the training and test sets.

largest deviation is −41.61 kJ mol$^{-1}$, and there are 70 outliers (i.e., where the error is larger than 8.37 kJ mol$^{-1}$ (2 kcal mol$^{-1}$)) among the 526 entries. In addition, we find that G4(MP2) performs well for all three subsets of E2.

When G4(MP2) is compared with G4(MP2)-6X (parametrized with E2), we can see that G4(MP2)-6X gives lower MAD and SD values overall, as well as for the individual subsets. The new procedure also produces fewer outliers for E2 and for the subsets. We find that G4(MP2)-6X performs only slightly better for the G2/97′ set, but there are major improvements for the E0 and E1 sets. For example, the MAD for the E0 set is improved by 1.15 kJ mol$^{-1}$, the SD is lowered by 2.11 kJ mol$^{-1}$, and the NO is reduced by 12. Thus, the inclusion of the six additional parameters improves the accuracy and precision for a wide range of systems, and these improvements come at effectively no additional computational cost. Notably, when trained by either the G2/97 or the E2 sets, G4(MP2)-6X achieves chemical accuracy with MADs of 3.85 kJ mol$^{-1}$ (0.92 kcal mol$^{-1}$) and 3.64 kJ mol$^{-1}$ (0.87 kcal mol$^{-1}$), respectively.

A comparison of the performance of the G2/97-parametrized G4(MP2)-6X with that parametrized using E2 reveals that the E2-parametrized method gives lower overall MAD, LD, SD, and NO values. When the statistics for the three component sets of E2 are examined, namely, E0, G2/97′, and E1, we find that the E1 set is the prime beneficiary of the parametrization with the E2 set, with notably reduced MAD and SD values. The E2-parametrization also gives better statistics for the E0 set, while the performance for the G2/97′ set is slightly worse, presumably due to the reduced weight of G2/97 properties in the E2 set.

While parametrization of G4(MP2)-6X with the E2 set leads to superior performance compared with fitting to G2/97, parametrization with the latter is not far behind, with an MAD and SD that are less than 0.5 kJ mol$^{-1}$ higher than the E2-parametrized values. Importantly, the G2/97 set appears to be very portable: using it for parametrization gives uniformly reasonable results for E0 (MAD = 3.81 kJ mol$^{-1}$) and E1 (3.44 kJ mol$^{-1}$) as well as for the G2/97′ set (4.09 kJ mol$^{-1}$). This demonstrates that G2/97 can be used as a cost-effective training set for preliminary screening in the formulation of composite procedures.

***Table 5.*** Deviations (kJ mol$^{-1}$) for the E2-Parameterized G4(MP2)-6X Procedure for the Various Test Sets[a]

| | MD | MAD | LD | SD | NO |
|---|---|---|---|---|---|
| E0 (148) | −0.42 | 3.43 | −19.47 | 4.67 | 8 |
| W4/08 (99) | −0.59 | 4.00 | −19.47 | 5.32 | 8 |
| DBH24 (24) | 1.39 | 2.97 | 7.30 | 3.47 | 0 |
| HB16 (16) | −1.69 | 1.78 | −5.17 | 1.59 | 0 |
| WI9/04 (9) | −1.23 | 1.34 | −2.31 | 0.88 | 0 |
| G2/97′ (248) | −0.49 | 4.14 | −36.66 | 5.87 | 29 |
| $\Delta H_f$ (94) | −0.65 | 3.04 | −16.62 | 4.61 | 6 |
| IE (88) | −0.39 | 4.68 | −36.66 | 6.75 | 12 |
| EA (58) | −0.30 | 5.29 | −15.53 | 6.56 | 11 |
| PA (8) | −1.12 | 2.79 | −7.99 | 3.74 | 0 |
| E1 (130) | 0.37 | 2.92 | 15.17 | 4.04 | 7 |
| G3/99′ (73) | 0.11 | 3.36 | 15.17 | 4.65 | 6 |
| ADD (21) | 1.80 | 2.47 | 7.21 | 2.95 | 0 |
| ABS (28) | 1.31 | 1.68 | 3.72 | 1.53 | 0 |
| PR8 (8) | −4.29 | 4.35 | −8.44 | 3.02 | 1 |
| E2 (526) | −0.26 | 3.64 | −36.66 | 5.14 | 44 |

[a] See section 3.3 for the definitions for the training and test sets.

When a procedure is developed by fitting a sizable number of parameters over a particular training set, one potential pitfall is that the resulting method may not be applicable to properties and systems that are not included in the training set. In this connection, our finding of comparable performance for the G4(MP2)-6X procedure when parametrized with the G2/97 and E2 training sets demonstrates to some extent the robustness of the form of parametrization in this method.

How does G4(MP2)-6X compare with G4? We have carried out such a comparison using the smaller E0 test set of 148 energies. The same BMK/6-31+G(2df,p) geometries were used for both G4(MP2)-6X and G4 for the sake of consistency.[35] We find that the G4(MP2)-6X procedure gives an MAD (3.43 kJ mol$^{-1}$) that is only slightly higher than that for G4 (3.22 kJ mol$^{-1}$) but considerably lower than that for G4(MP2) with BMK geometries (4.03 kJ mol$^{-1}$).

We now look at the performance of the E2-parametrized G4(MP2)-6X procedure in more detail (Table 5). We find that G4(MP2)-6X performs well for AEs and $\Delta H_f$s (W4/08, G2/97′ $\Delta H_f$, and G3/99′), PAs, radical reaction enthalpies (ADD and ABS), intermolecular interactions (HB16 and WI9/04), and barriers in the DBH24 set, with MADs that are in all cases smaller than 4.18 kJ mol$^{-1}$ (1 kcal mol$^{-1}$). On the other hand, pericyclic reaction barriers (PR8), IEs, and EAs show larger MAD values.

**4.4. Outliers in the E2 Set.** Table 6 lists the outliers in the E2 set, i.e., situations for which the G4(MP2)-6X energy deviates from the corresponding benchmark value by more than 8.37 kJ mol$^{-1}$ (2 kcal mol$^{-1}$). Out of the 44 outliers, there are four cases where the deviation is larger than 16.74 kJ mol$^{-1}$ (4 kcal mol$^{-1}$). These are the AEs for $ClOO^{\bullet}$ and $P_4$, and the IEs for $B_2F_4$ and $B_2H_4$. In addition, there are 10

***Table 6.*** Outliers for the E2 Set for the G4(MP2)-6X Procedure Where Deviations ($D$) Are Larger than 8.37 kJ mol$^{-1}$ (2 kcal mol$^{-1}$)[a]

| | | Range of absolute deviations ($D$, kcal mol$^{-1}$) | | |
|---|---|---|---|---|
| | | $2 < D < 3$ | $3 < D < 4$ | $4 < D$ |
| W4/08 | AE | $BHF_2$, $F_2O_2$, $CS_2$, $S_4$, CS | $B_2H_6$ | ClOO, $P_4$ |
| DBH24 | $\Delta H^{\ddagger}$ | | nil | |
| HB16 | HB | | nil | |
| WI9/04 | BE | | nil | |
| G2/97′ | $\Delta H_f$ | $SiH_2$ singlet, thiophene | $CF_2O$, $C_2F_4$, $C_2Cl_4$, $CH_2CHCl$ | |
| G2/97 | IE | S, $CH_4$, $BF_3$, $BCl_3$, $C_6H_6$, $C_6H_5CH_3$, $CH_3F$, $C_6H_5NH_2$, $Si_2H_6$ | $C_3H_7$ | $B_2F_4$, $B_2H_4$ |
| G2/97 | EA | Li, F, Si, P, CH, SiH, PH, CCH, $CH_2NC$ | B, C | |
| G2/97 | PA | | nil | |
| G3/99′ | $\Delta H_f$ | benzoquinone, pyrimidine, 2-methylthiophene, $SF_6$ | acetylacetylene, tetramethylsilane | |
| RR49 | $\Delta H_r$ | | nil | |
| PR8 | $\Delta H^{\ddagger}$ | 1,3-cyclopentadiene + $C_2H_4$ → norbornene | | |

[a] See section 3.3 for the definitions for the training and test sets; AE = atomization energy, $\Delta H^{\ddagger}$ = barrier, HB = hydrogen-bond energy, BE = binding energy, $\Delta H_f$ = heat of formation, IE = ionization energy, EA = electron affinity, PA = proton affinity, $\Delta H_r$ = reaction energy.

***Table 7.*** Comparison of Deviations (kJ mol$^{-1}$) for Various Alternative Procedures for the Outliers Listed in Table 6 Where the Deviations Are Larger than 12.55 kJ mol$^{-1}$ (3 kcal mol$^{-1}$)[a]

| | | G4(MP2)-6X | G4(MP2) | G4 | W1U | W1Usc |
|---|---|---|---|---|---|---|
| AE | ClOO | −19.47 | −18.51 | −16.86 | −18.53 | −2.99 |
| AE | $P_4$ | −16.83 | 9.84 | −2.66 | −4.30 | −4.34 |
| IE | $B_2F_4$ | −36.66 | −41.61 | −37.27 | −32.52 | −32.59 |
| IE | $B_2H_4$ | −16.81 | −9.29 | −9.50 | −12.28 | −12.37 |
| AE | $B_2H_6$ | −15.98 | −17.51 | −3.62 | 3.42 | 3.42 |
| $\Delta H_f$ | $CF_2O$ | 16.53 | 17.67 | 16.39 | 8.76 | 9.22 |
| $\Delta H_f$ | $C_2F_4$ | −12.63 | −12.51 | −13.50 | −28.49 | −27.88 |
| $\Delta H_f$ | $C_2Cl_4$ | −13.89 | −23.39 | −13.12 | −22.94 | −21.88 |
| $\Delta H_f$ | $CH_2CHCl$ | −16.62 | −19.11 | −15.17 | −22.08 | −21.56 |
| IE | $C_3H_7$ | 13.95 | 14.02 | 12.17 | 7.48 | 7.72 |
| EA | B | −15.53 | −21.44 | −9.15 | −7.96 | −7.80 |
| EA | C | −14.78 | −18.85 | −9.86 | −6.56 | −6.62 |
| $\Delta H_f$ | acetylacetylene | 13.45 | 11.04 | 10.68 | 6.72 | 7.56 |
| $\Delta H_f$ | tetramethylsilane | 15.17 | 15.13 | 14.36 | 0.72 | 1.66 |

[a] AE = atomization energy, $\Delta H_f$ = heat of formation, IE = ionization energy, EA = electron affinity.

cases with a deviation that falls between 12.55 kJ mol$^{-1}$ (3 kcal mol$^{-1}$) and 16.74 kJ mol$^{-1}$. We have briefly examined these cases using higher-level procedures. The results are shown in Table 7.

When comparing G4(MP2) and G4(MP2)-6X with G4, we generally find better agreement between G4(MP2)-6X and G4 than between G4(MP2) and G4. However, there are a number of cases where G4(MP2) agrees better with G4, notably for the IE of $B_2H_4$.

We find that G4 significantly reduces the deviations from benchmark values for $P_4$ (AE), $B_2H_4$ (IE), $B_2H_6$ (AE), B (EA), and C (EA). Presumably, the use of MP4 in G4 (instead of MP2 in G4(MP2) and G4(MP2)-6X) to account for the effect of adding diffuse and polarization functions to the 6-31G(d) basis set is beneficial in these cases. The use of the W1U procedure[36] further reduces the number of outliers with deviations that are larger than 12.55 kJ mol$^{-1}$ from seven to five. However, we notice that the use of W1U leads to notably larger deviations than with G4 in a number of cases, particularly for the $\Delta H_f$ of $C_2F_4$, $C_2Cl_4$, and $CH_2CHCl$. The use of the spin correction (sc) term in the W1Usc procedure reduces the deviation for the AE of ClOO substantially compared with W1U (from −18.53 kJ mol$^{-1}$ to −2.99 kJ mol$^{-1}$), but in general the two procedures perform comparably.

For the four cases where the deviations at the W1Usc level remain larger than 12.55 kJ mol$^{-1}$, namely, the IE of $B_2F_4$ and the $\Delta H_f$ of $C_2F_4$, $C_2Cl_4$, and $CH_2CHCl$, it has previously been suggested, on the basis of calculations of isodesmic reactions and G3 and G4 calculations, that the accuracy of the experimental values is questionable.[28] The large deviations for W1U and W1Usc in these cases support the desirability of experimental re-examination.

## 5. Concluding Remarks

We have developed the G4(MP2)-6X procedure as a variant of G4(MP2). It employs BMK/6-31+G(2df,p) for geometry optimization and has six additional scaling factors for the correlation energy components compared with G4(MP2), as well as reoptimized HLC parameters.

We find it to be an improvement (MAD = 3.64 kJ mol$^{-1}$) over G4(MP2) (MAD = 4.42 kJ mol$^{-1}$) for a wide range of properties for the E2 set of 526 energies. A comparison of G4(MP2)-6X with G4 on the E0 set shows that the performance of the new procedure (3.43 kJ mol$^{-1}$) approaches that for G4 (3.22 kJ mol$^{-1}$) for this series of energies, for which G4(MP2) has an MAD of 4.03 kJ mol$^{-1}$. Thus, the G4(MP2)-6X procedure bridges the gap between G4(MP2) and G4, with an accuracy that is closer to G4 but a computational cost that is comparable to that for the considerably more economical G4(MP2). While G4(MP2)-6X represents a noticeable improvement over G4(MP2), it also has cases of dramatic failures. These can generally be attributed to inadequate treatment of electron correlation.

We find that, when the G2/97 training set is used (in place of E2) for parametrization, G4(MP2)-6X still gives uniformly good results, not only for the types of energies represented in G2/97 but also for the wider range of energies represented in the larger E2 set. This reflects to some extent the robustness of G4(MP2)-6X, despite the additional parameters. In addition, it suggests the more general use of G2/97 as a cost-effective training set for the screening of theoretical procedures.

**Acknowledgment.** We gratefully acknowledge the award of an Australian Professorial Fellowship and funding from the ARC Centre of Excellence for Free Radical Chemistry and Biotechnology (to L.R.) and generous alloca-

tions of computer time from the National Computational Infrastructure (NCI) National Facility and Intersect.

**Supporting Information Available:** Script for running G4(MP2)-6X calculations with Gaussian 09; reactions of the DBH24, RR49, and PR8 sets (Table S1); zero-point vibrational energies and thermal corrections from scaled BMK/6-31+G(2df,p) frequencies and G4(MP2)-6X total electronic energies (Table S2); and deviations from experimental and benchmark values (Table S3). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) For general overviews, see: (a) Helgaker, T.; Klopper, W.; Tew, D. P. *Mol. Phys.* **2008**, *106*, 2107. (b) Martin, J. M. L. *Annu. Rep. Comput. Chem.* **2005**, *1*, 31. (c) Raghavachari, K.; Curtiss, L. A. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005.

(2) (a) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622. (b) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221. (c) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764. (d) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108.

(3) (a) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A, Jr. *J. Chem. Phys.* **1996**, *104*, 2598. (b) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822. (c) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A *J. Chem. Phys.* **2000**, *112*, 6532. (d) Wood, G. P. F.; Radom, L.; Petersson, G. A.; Barnes, E. C.; Frisch, M. J; Montgomery, J. A, Jr. *J. Chem. Phys.* **2006**, *125*, 094106.

(4) (a) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843. (b) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129. (c) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.

(5) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.

(6) (a) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 1293. (b) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703. (c) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105.

(7) See, for example: (a) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986. (b) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley: New York, 2001. (c) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, U. K., 2007.

(8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(9) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.

(10) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklaâ, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO 2006.1*; University of Birmingham: Birmingham, U.K., 2006.

(11) Crawford, T. D; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610.

(12) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.

(13) Merrick, J. P.; Moran, D.; Radom, L. *J. Phys. Chem. A* **2007**, *111*, 11683.

(14) The scale factors for the 6-31+G(2df,p) basis set are taken as the average of the 6-31G(2df,p) and 6-311+G(2df,p) values obtained from ref 13, namely, 0.9770 for the ZPVE and 0.9627 for $\Delta H$. For example, the BMK/6-31+G(2df,p) scale factor for the ZPVE at 298 K is calculated as $1/2 \times [0.9752$ (BMK/6-31G(2df,p)) + 0.9787 (BMK/6-311+G(2df,p))] = 0.9770.

(15) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 1125.

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

(17) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(18) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2008**, *128*, 124111.

(19) We have explored the effect of re-optimizing the exponent for the extrapolation procedure and found that this does not lead to notable improvements.

(20) (a) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569. (b) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808.

(21) Boese, A. D.; Martin, J. M. L.; Klopper, W. *J. Phys. Chem. A* **2007**, *111*, 11122.

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

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

(24) (a) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063. (b) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 42.

(25) Karton, A.; Tarnopolsky, A.; Lamère, J.-F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868.

(26) Berkowitz, J.; Chupka, W. A.; Walter, T. A. *J. Chem. Phys.* **1969**, *50*, 1497.

(27) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014.

(28) Curtiss, L.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(29) Lin, C. Y.; Hodgson, J. L.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2009**, *113*, 3690.

(30) (a) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. *J. Phys. Chem. A* **2003**, *107*, 11445. (b) Ess, D. H.; Houk, K. N. *J. Phys. Chem. A* **2005**, *109*, 9542.

(31) Chan, B.; Coote, M. L.; Radom, L. *J. Chem. Theory Comput.* **2010**, *6*, 2647.

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

(33) Zheng, W.-R.; Fu, Y.; Guo, Q.-X. *J. Chem. Theory Comput.* **2008**, *4*, 1324.

(34) We also find that, by optimizing both G4(MP2)//B3-LYP/6-31G(2df,p) and G4(MP2)//BMK/6-31+G(2df,p) with the E2 set, the BMK-optimized procedure leads to a comparable MAD (4.18 kJ $mol^{-1}$) to that for the B3-LYP-optimized procedure (4.25 kJ $mol^{-1}$). The BMK-optimized procedure has a lower SD (5.88 kJ $mol^{-1}$) than that for the B3-LYP-optimized procedure (6.20 kJ $mol^{-1}$), as well as a lower NO (68, compared with 71 for G4(MP2)//B3-LYP/6-31G(2df,p)).

(35) When the default B3-LYP geometries in G4 and G4(MP2) were used, the MADs for the E0 set were 3.36 and 4.58 kJ $mol^{-1}$, respectively, for G4 and G4(MP2).

(36) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A.; Frisch, M. J.; Martin, J. M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2687.

CT100542X

JCTC Journal of Chemical Theory and Computation

# Bonding Conundrums in the $C_2$ Molecule: A Valence Bond Study

Peifeng Su,[†] Jifang Wu,[†] Junjing Gu,[†] Wei Wu,*[,†] Sason Shaik,*[,‡] and Philippe C. Hiberty*[,§]

*The State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, Fujian 361005, China, Institute of Chemistry and The Lise Meitner-Minerva Center for Computational Quantum Chemistry, The Hebrew University, Jerusalem, 91904, Israel, and Laboratoire de Chimie Physique, Groupe de Chimie Théorique, CNRS UMR 8000, Université de Paris-Sud, 91405 Orsay Cédex, France*

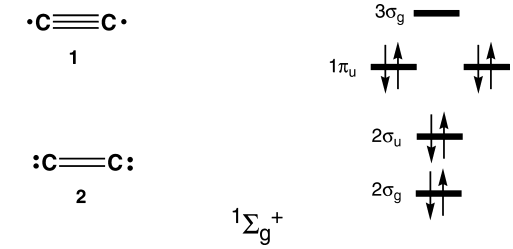
Received October 7, 2010

**Abstract:** The ab initio VB study for the electronic structure of the $C_2$ molecule in the ground state is presented in this work. VB calculations involving 78 chemically relevant VB structures can predict the bonding energy of $C_2$ quite well. Sequentially, a VBCIS calculation provides spectroscopic parameters that are very close to full CI calculated values in the same basis set. Furthermore, the analysis of the bonding scheme shows that a triply bonded structure is the major one in terms of weights, and the lowest in energy at the equilibrium distance. The second structure in terms of weights is an ethylene-like structure, displaying a $\sigma + \pi$ double bond. The structure with two suspended $\pi$ bonds but no $\sigma$ bond contributes only marginally to the ground state. This ordering of weights for the VB structures describing the $C_2$ molecule is shown to be consistent with the shape of the molecular orbitals and with the multireference character of the ground state. With the triply bonded bonding scheme, the natures of the $\pi$ and $\sigma$ bonds are investigated, and then the corresponding "in situ" bond strengths are estimated. The contribution of the covalent-ionic resonance energy to $\pi$ and $\sigma$ bonding is revealed and discussed.

## Introduction

The precise bonding in the $C_2$ molecule is quite enigmatic if not intriguing. Applying the principle of maximum coupling between overlapping atomic orbitals would lead to structure **1**, where the triple bond is made from one $\sigma$ and two $\pi$ bonds as in acetylene, and the two odd electrons on the carbon atoms are singlet coupled (Scheme 1). On the other hand, when a molecular orbital (MO) diagram is used,

***Scheme 1.*** Schematic Bonding Schemes and MO Diagram for the $X^1\Sigma_g^+$ Ground State of $C_2$




in Scheme 1, and the simple formula for calculating bond order is applied, one would predict that the ground state, $X^1\Sigma_g^+$, is a doubly bonded molecule, as depicted in the simple valence bond (VB) cartoon in **2**. From the MO diagram, the double bond in **2** corresponds to two suspended $\pi$ bonds, whereas the two "lone pairs" on the carbon atoms

* To whom correspondence should be addressed. Tel.: +86-5922182825 (W.W.), +972-26585909 (S.S.), +33-0169156175 (P.C.H.). Fax: +86-5922184708 (W.W.), +972-26584680 (S.S.), +33-0169154447 (P.C.H.). E-mail: weiwu@xmu.edu.cn (W.W.), sason@yfaat.ch.huji.ac.il (S.S.), philippe.hiberty@u-psud.fr (P.C.H.).

[†] Xiamen University.

[‡] The Hebrew University.

[§] Université de Paris-Sud.

10.1021/ct100577v 
Published on Web 12/20/2010

***Table 1.*** The Lowest Seven Electronic States of $C_2$ Molecule[a]

| term | main electronic configuration | $\Delta E$ ($cm^{-1}$) |
|---|---|---|
| $X^1\Sigma_g^+$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^4$ | |
| $a^3\Pi_u$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^3(\sigma_g2p)$ | 716 |
| $b^3\Sigma_g^-$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^2(\sigma_g2p)^2$ | 6434 |
| $A^1\Pi_u$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^3(\sigma_g2p)$ | 8391 |
| $c^3\Sigma_u^+$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)(\sigma_g2p)^2(\pi_g2p)$ | 9227 |
| $B^1\Delta_g$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^2(\sigma_g2p)^2$ | ~11858[b] |
| $B'^1\Sigma_g^+$ | $KK(\sigma_g2s)^2(\sigma_u2s)^2(\pi_u2p)^2(\sigma_g2p)^2$ | ~15196[b] |

[a] Experimental values taken from the literature.[17,19] [b] Estimated from Figure 1 in ref 17.

are $\sigma$ lone pairs, which result from the full occupancy of the bonding and antibonding MO pair $2\sigma_g$ and $2\sigma_u$.

However, as pointed out recently,[1] the C−C bond length of 1.243 Å in $X^1\Sigma_g^+$ [2] is shorter than any known C═C double bond, e.g., in ethylene.[3−5] Thus, Jemmis et al.[1] and others[6] argued that suspended $\pi$ bonds are in fact shorter than $\sigma$ bonds, and as a result, the $\pi$ doubly bonded molecule has a shorter C−C distance than, e.g., ethylene with a $\sigma + \pi$ double bond. Nevertheless, this assessment of the bonding in $C_2$ rests on the supposed antibonding nature of the doubly occupied $2\sigma_u$ MO, which is assumed to compensate for the bonding nature of $2\sigma_g$ and to result in the absence of any $\sigma$ bonding. However, an alternative characterization of the $2\sigma_g-2\sigma_u$ MO pair, in which the $2\sigma_u$ MO would be only weakly antibonding and would not compensate for the strongly bonding character of $2\sigma_g$, would make the C−C bonding approach a triple bond,[7,8] in support of structure **1** rather than **2**. The respective shapes of the $2\sigma_g$ and $2\sigma_u$ MOs are thus of fundamental importance for the C−C bonding assessment and will be examined below.

Another intriguing aspect of the $C_2$ molecule is the presence of a low-lying triplet state (Table 1), $a^3\Pi_u$, with a main configuration $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^3 3\sigma_g^1\rangle$, lying only 2 kcal/mol above the ground state $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^4\rangle$. Especially interesting is the fact that the bond length of this excited state, 1.313 Å, is significantly increased relative to the ground state, despite the quasi-identical bonding energy of the two states. Qualitatively, this breaking of the bond-length/bonding-energy relationship might be interpreted in two ways: (i) Assuming structure **2** for the ground state, as done by Sherrill et al.,[9−11] the $X^1\Sigma_g^+ \rightarrow a^3\Pi_u$ weakens one $\pi$ bond and benefits the $\sigma$ bond; in such a case, the corresponding elongation is expected according to the author's view that suspended $\pi$ bonds should be shorter than $\sigma + \pi$ bonds.[9−11] (ii) If, on the other hand, the triply bonded structure **1** is assumed, the excitation cannot significantly reinforce the $\sigma$ bond, which is already present in the ground state but may transform a two-electron $\pi$ bond into a one-electron $\pi$ bond. One-electron bonds are not necessarily weaker than two-electron bonds when the interatomic orbital overlap is weak (see, e.g., $Li_2^+$ vs $Li_2$),[12] but they are always longer, thus explaining the long C−C bond in the $a^3\Pi_u$ state.

Still, since the ground state has a longer C−C bond than acetylene, it is clear that structure **1** alone cannot account for the bonding in $C_2$, unless the $\sigma$ bond is particularly weak, and the question is whether one should regard the molecule more as a hybrid of **1** and **2** or simply as **1** with a weak $\sigma$ bond, or perhaps having different bonding features altogether? This, as well as other questions about $C_2$, can be answered by modern VB calculations,[13] which are presented in this work to address a host of bonding conundrums exhibited by this diatomic molecule.

Interestingly, despite the small size of the molecule, it has proven to be a "hard nut to crack" by theoretical means, and therefore before addressing the bonding issues, it is appropriate to summarize some experimental facts and discuss some selected previous theoretical studies on $C_2$. The molecule is present in astrophysical environments and many chemical processes in the gas phase.[14−19] It has been observed during the photodissociation of acetylene and can be formed by the direct reaction of the $C(^3P)$ atom with CH.[16,20] Since the 1950s, many experimental and theoretical studies have been dedicated to the study of the features of $C_2$.[14−23] Recently, 17 electronic states of $C_2$ have been characterized by experimental methods. Using the orbital diagram in Scheme 1, Table 1 shows the electronic configurations and experimentally measured relative energies of the lowest seven electronic states.[17,19] As already pointed out, the ground state of the molecule is $X^1\Sigma_g^+$, and above it there are a few triplet and singlet states within 2−44 kcal/mol. The spectroscopic parameters and potential energy curves of the ground state and some of the low lying excited states have been studied since 1992 using high-level theoretical methods. Bartlett and Watts[21] studied the $X^1\Sigma_g^+$, $a^3\Pi_u$, and $b^3\Sigma_g^-$ states and found that the single reference CCSD method cannot describe the $C_2$ molecule qualitatively due to the multireference character of its ground state. Pradhan et al.[22] used the multireference IC-MRCI method and obtained satisfactory quantitative results for the spectroscopic parameters in the ground state. Subsequently, Halvick[20] used the EHFACE2 (extended Hartree−Fock approximate correlation energy) model[24,25] parameters fitted from ab initio MRCI calculations in the large correlation-consistent cc-pV5Z basis set and calculated the adiabatic potential energy curves of the 12 lowest electronic states of the molecule. Sherrill and Abrams[9] studied the potential energy curves of the $X^1\Sigma_g^+$, $B^1\Delta_g$, and $B'^1\Sigma_g^+$ states at the full CI/6-31G* level and used this level as a benchmark for the methods that are based on a single reference molecular orbital wave function. These authors concluded that electron correlation methods based on an UHF (Unrestricted Hartree−Fock) single reference can describe the $C_2$ molecule correctly, whereas the methods that are based on the RHF (Restricted Hartree−Fock) reference are wrong. Recently, Sherrill et al.[10] evaluated the performance of multireference methods such as CASSCF, CASPT2, and MRCI and renormalized single reference methods such as EOMCCSD and CR-EOMCCSD(T).[11] The authors concluded that the multireference methods give results on par with full CI. Moreover, the renormalized EOMCCSD and CR-EOMCCSD(T) methods have much better performance compared with the straightforward single reference methods CCSD and CCSD(T). Very recently, Varandas extrapolated MRCI results to complete basis set limits,[26] and Mahapatra et al. tested the state-specific multireference perturbation theory approach.[27] A very accurate binding energy for the $C_2$ ground state was also obtained by Bytautas and Ruedenberg,[28] who approximated full CI results with the "correlation energy extrapolation by intrinsic scaling"

(CEEIS) method in double-, triple-, and quadruple-ζ basis sets with extrapolation to the complete basis set limit.

What emerges from all of these high level studies is the extreme difficulty in calculating the ground state and the low-lying excited states of $C_2$ in a meaningful way, owing to the multireference character of the wave functions and the near degeneracies which change very rapidly as a function of the C−C distance. As noted by Sherrill and Piecuch,[11] the low-lying states of $C_2$ are so challenging and the failures of various high-level electronic structure methods are so dramatic that it is desirable to assess the reliability of the methods against full CI results.

The multireference character of the ground state of $C_2$ is apparent already at the equilibrium geometry, wherein the coefficient of the primary configuration, $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^4\rangle$, is only 0.83 in a full CI calculation,[11] and the doubly excited configuration $|(\text{core})2\sigma_g^2 1\pi_u^4 3\sigma_g^2\rangle$ has a surprisingly large coefficient of 0.33. The situation complicates further near 1.7 Å, where two diexcited determinants of the type $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^2 3\sigma_g^2\rangle$, which have six electrons in $\sigma$ orbitals and only two in a $\pi$ orbital, come into play and become dominant in the $X^1\Sigma_g^+$ state due to avoided crossing with the $B'^1\Sigma_g^+$ state.[11]

In summary, the studies using sophisticated theoretical methods show that the ground state has distinct multireference characteristics.[9−11] However, the simple questions posed at the outset (Scheme 1) regarding the type and number of bonds that account for the bond energy and bond length in the molecule are not discussed deeply in their work. Does $C_2$ have a suspended double $\pi$-bonding as in **2** or a triple bond as in **1**, or even another combination of $\sigma$ and $\pi$ bonds, and if so, what is the precise role of the $\sigma$ bonding? These questions concerning chemical bonding will be addressed by use of the VB method, which has both multireference character and conceptual clarity suitable to answer such questions.

The article is organized as follows: It begins with the introduction of the ab initio VB methods. Second, the VB structures that are involved in the calculation are selected. Third, the computational details and the results are shown and discussed. The article is ended with a brief conclusion.

## Theory and Methodology

A many-electron system wave function Ψ in VB theory is expressed as a linear combination of Heitler−London−Slater−Pauling (HLSP) functions, $\Phi_K$, in eq 1:[29]

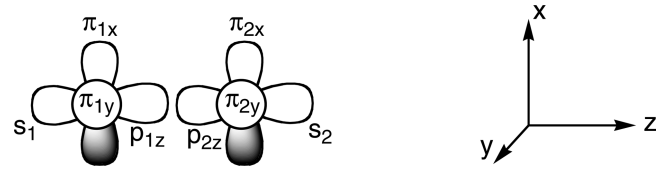
$$\Psi = \sum_K C_K \Phi_K \tag{1}$$

where $\Phi_K$ corresponds to "classical" VB structures and $C_K$ represents structural coefficients.

There are several computational approaches for VB theory at the ab initio level. In the VBSCF procedure,[30,31] both the VB orbitals and structural coefficients are optimized simultaneously to minimize the total energy. As such, the VBSCF method takes care of the static electron correlation, but it lacks dynamic correlation.

The VBCI method,[32,33] which uses a configuration interaction technique, following VBSCF calculation, considers the dynamic correlation by involving excited VB structures which are generated by replacing occupied orbitals with virtual orbitals. The virtual orbitals are strictly localized on precisely the same fragment as the corresponding occupied orbitals. In this manner, by merging all of the excited VB structures into the corresponding fundamental structures of the same electron occupancy, the VBCI wave function is condensed to a linear combination of the same minimal number of VB structures as in the VBSCF method. The VBCI computations are performed at the VBCIS level, which involves single excitations only.

***Scheme 2.*** The VB-Orbital Representation in a Coordinate Axis[a]




[a] The $p_y$ orbitals are drawn with one lobe pointing at the observer.

The weights of the VB structures can be defined in several ways. One commonly used definition is the Coulson−Chirgwin formula,[34] eq 2, which is the equivalent of a Mulliken population analysis in VB theory.

$$W_K^{\text{Coulson}} = C_K^2 + \sum_{L \neq K} C_K C_L \langle \Phi_K | \Phi_L \rangle \tag{2}$$

One drawback of this formula is that the second term may become more important than the first one if the overlap between VB structures is too large, possibly leading to some negative weights. Such weights, which are nonphysical, are generally interpreted as an indication that the VB structure in question is unimportant; however, eq 2 becomes inappropriate when negative weights get large absolute values. For this reason, other definitions have been proposed, among which are the Löwdin weights[35] in eq 3:

$$W_K^{\text{Lowdin}} = \sum_{I,J} S_{KI}^{1/2} C_I S_{KJ}^{1/2} C_J \tag{3}$$

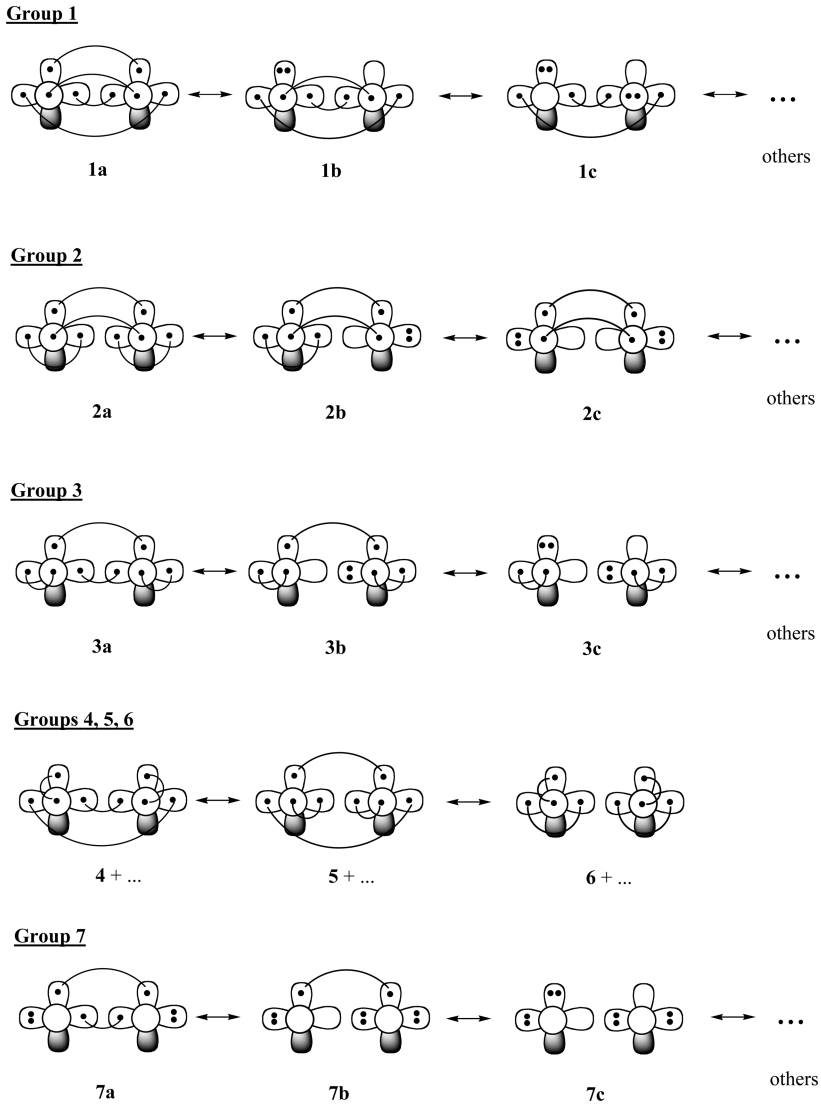
or the renormalized sum of the coefficients squares in eq 4 where $N$ is the normalizing factor.

$$W_K^{\text{renorm}} = N C_K^2;\ N = \sum_K \frac{1}{C_K^2} \tag{4}$$

Throughout the work, we use eqs 2−4, except for large distances (2.0 Å), where the Chirgwin−Coulson definition was abandoned since it gave large negative weights.

**Computational Details.** All eight valence electrons of the $C_2$ molecule are involved in VB calculations. As shown in Scheme 2, the two atoms lie on the $z$ axis, and the $2p_z$ and 2s atomic orbitals (AOs) of the C atoms are hybridized to form two hybrid orbitals pointing toward each other, labeled $p_{1z}$ and $p_{2z}$, and other hybrid orbitals pointing outward, labeled $s_1$ and $s_2$. The remaining AOs are two pure p orbitals lying in the $xz$ plane, $\pi_{1x}$ and $\pi_{2x}$, and two other pure p AOs

***Scheme 3.*** Some Representative VB Structures Gathered by Groups of Bonding Modes




in the $yz$ plane, $\pi_{1y}$ and $\pi_{2y}$, drawn as circles with one lobe pointing at the observer.

The VBSCF and VBCIS methods are used for the calculations of spectroscopic parameters, with the 6-31G* and cc-pVTZ basis sets. The diabatic and adiabatic potential energy curves are calculated at the VBSCF and CASSCF(8,8) levels in the 6-31G* basis set. All orbitals are strictly localized; i.e., they are expressed as combinations of basis functions that belong to single carbon atoms, without tails on the other atom. The VB calculations are carried out with the Xiamen Valence Bond (XMVB) package of programs.[36,37] Basis set integral and nuclear repulsion energy are taken from the output of Gaussian 98[38] calculations.

**The VB Structure Set.** For a system of spin $S$ with $N$ electrons and $m$ orbitals, the number of independent VB structures is given by the Weyl formula:[39]

$$D(m,N,S) = \frac{2S+1}{m+1}\binom{m+1}{\frac{1}{2}N+S+1}\binom{m+1}{\frac{1}{2}N-S} \quad (5)$$

For the singlet ground state of $C_2$, taking all eight valence orbitals and electrons into account, the total number of VB canonical structures amounts to 1764. Of course, not all of these VB structures are essential for the description of the bonding in $C_2$, and the first step of the VB application is to select the VB structures that are necessary and sufficient for a reliable description of the electronic state in question. Clearly, an effective way is to select the VB structures by analyzing the characteristics of chemical bonding. Scheme 3 displays some generic VB structures which are gathered by groups, each group representing a specific bonding mode. The complete set of VB structures is displayed in the Supporting Information.

Group 1 involves the VB structures needed to represent the bonding mode with two $\pi$ bonds, one $\sigma$ bond, and two unpaired electrons in the $s_1$ and $s_2$ orbitals, i.e., structure **1** in Scheme 1. Since each bond is a combination of a major covalent component and two relatively minor ionic components, we must take all combinations of covalent and ionic VB structures. Thus, in Scheme 3, **1a** represents the fully covalent triply bonded VB structure, while **1b** is one of the monoionic structures. The dionic structures are also included, but only if they keep the two carbon atoms neutral, as in **1c**. Triply ionic structures are neglected. As a result, a total of 21 VB structures are kept for the description of the triply bonded structure **1**.

The same principles are applied to select the VB structures needed for the description of structure **2** of Scheme 1, which possesses two suspended $\pi$ bonds. A total of 29 structures are selected (see the Supporting Information) and gathered in group 2, among which, three representative structures are **2a**−**c** in Scheme 3.

Another doubly bonded structure is possible with $\sigma + \pi$ bonds, while the electrons of the remaining $\pi$ system are relegated to orbitals $s_1$ and $s_2$, as shown in **3a**−**c** in Scheme 3. Group **3** involves a total of 14 VB structures.

To complete the set of VB structures having four electrons in $\sigma$ and four electrons in $\pi$ orbitals, and to ensure correct dissociation of the $C_2$ molecule, one must make sure that all 14 ways of spin-coupling the eight electrons wherein all orbitals are singly occupied (e.g., **1a**, **2a**, **3a**, and so on) are included in the VB structure set. Such structures are gathered in groups 4, 5, and 6 in Scheme 3a. Two VB structures (group 4) display a single $\sigma$ bond, four structures display a single $\pi$ bond (group 5), and eight structures display no bond at all (group 6). Accordingly, the total number of VB structures of the 4-$\sigma$−4-$\pi$ type amounts to 78 and constitutes what we will call hereafter the 78-set.

To ensure a correct dissociation, the description of $C_2$ requires also the VB structures displaying six electrons in $\sigma$ orbitals and only two in $\pi$ orbitals. According to full CI benchmarks, such structures are expected to play little or no role at the equilibrium distance but should become important near 1.7 Å.[9−11] Three of these 6-$\sigma$−2-$\pi$ structures, out of a total of 14 selected ones, are represented in group 7 in Scheme 3. These 14 VB structures, added to the 78-set, form the 92-set that will be used to generate the ground state potential energy curve (PEC) from the equilibrium to a separation of 3.0 Å.

Hereafter, we shall use the term "group" to designate either a set of VB structures or the bonding scheme that these structures represent. For example, "group 1" or "structure **1**" will designate the triply bonding structure displayed in Scheme 1 as well as the combination of the 21 VB structures that are needed to represent it, and so on.

**Calculating the Energies of the Individual Group Structures.** The energy of an individual group of structures can be calculated by block-diagonalizing a Hamiltonian matrix involving only the VB structures that belong to this group, i.e., VB structures 1−21 for group 1, 22−50 for group 2, and so on. This can be done by keeping the same orbitals as those optimized for the ground state, without further orbital reoptimization. This technique is appropriate if the aim is to better analyze the properties of the ground state in terms of VB mixing. On the other hand, one may also perform a full VB calculation of the group structure alone, this time by reoptimizing the orbitals. With this latter method, called the "variational diabatic configuration (VDC) method", the energy of the group structure is variationally minimized. Both methods have been used in this work.

## Computational Results and Discussion

**Dissociation Energy Curve and Spectroscopic Constants for the $X^1\Sigma_g^+$ Ground State.** Before discussing the bonding mode of $C_2$, it is important to check that our 92-set of VB structures is sufficient to faithfully describe the electronic structure of this molecule at any distance. This can be done by plotting the adiabatic PEC for the $X^1\Sigma_g^+$ state as a function of the interatomic distance and comparing it with larger VB or CASSCF results. Figure 1 shows the dissociation energy curve for the $X^1\Sigma_g^+$ state as calculated at the VBSCF level with the 92-set, at the full valence VBSCF level involving 1764 spin-couplings, and at the full valence CASSCF(8,8) level in the MO framework. It can be seen that the VBSCF-92 curve is very close to the VBSCF-1764 one at all distances, especially at the equilibrium distance. The VBSCF-1764 curve is also quite close to the CASSCF(8,8) curve at the equilibrium and at large distances but somewhat departs from it in the region near 2 Å. This slight difference between the full valence VBSCF and CASSCF calculations might appear surprising, given that the two calculations span the same space of 1764 configurations. This is due to the fact that (i) the VBSCF-1764 calculation uses the orbitals of the VBSCF-92 calculation, without further reoptimization, and (ii) orbital optimization in CASSCF has a few more degrees of freedom than in VBSCF, as has been noted,[13] because the atomic orbitals (AOs) that compose the various MOs in CASSCF may be different in size and shape from one MO to another, while the set of AOs is unique in VBSCF. Of course, this VBSCF limitation disappears when further CI is performed, as in the VBCIS level. We note that the VBSCF/CASSCF difference is small, and both calculations yield the same bonding energies to within 6% using the 6-31G* basis set.

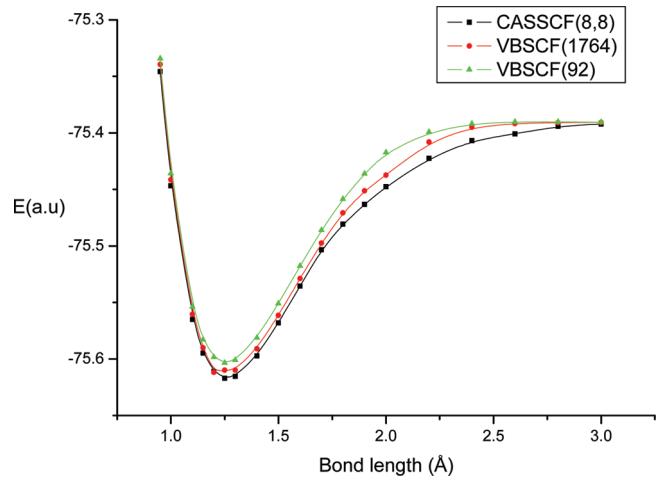



***Figure 1.*** The PECs of the ground state of the $C_2$ molecule computed at the VBSCF level in the 6-31G* basis set.

Table 2 shows the spectroscopic constants as computed by VBSCF and VBCIS methods for the ground state of $C_2$ molecule. These constants have been calculated using the restricted set of 78 VB structures, as the 6-$\sigma$−2-$\pi$ structures proved completely negligible at the equilibrium C−C distance. A comparison is made with experimental results and with some previous theoretical calculations. The VB optimized equilibrium bond lengths range between 1.252 and 1.262 Å, very close to the full CI value[9−11] (1.260 Å) and to other computational results by high level MO methods.[21,22] The VBSCF values of dissociation energy, 5.77 eV (6-31G*) and 5.80 eV (cc-pVTZ), amount to 93% of the experimental

***Table 2.*** The Spectroscopic Parameters of the Ground State of $C_2$

| method | basis set | $R_{eq}$ (Å) | $D_e$ (eV) | $\omega_e$ ($cm^{-1}$) |
|---|---|---|---|---|
| VBSCF | 6-31G* | 1.256 | 5.77 | 1994 |
| | cc-pVTZ | 1.252 | 5.80 | 1996 |
| VBCIS | 6-31G* | 1.262 | 6.29 | 1922 |
| | cc-pVTZ | 1.258 | 6.38 | 1895 |
| CASSCF(8,8) | 6-31G* | 1.262 | 6.15 | 1858 |
| | cc-pVTZ | 1.256 | 6.18 | 1840 |
| ICMRCI[a] | cc-pVTZ | 1.252 | 6.09 | 1841 |
| ICMRCI+Q[a] | cc-pVTZ | 1.253 | 6.01 | 1840 |
| CCSD(T)[b] | cc-pVTZ | 1.245 | 6.21 | 1869 |
| full CI[c] | 6-31G* | 1.260 | 6.00 | 1859 |
| exptl.[d] | | 1.243 | 6.42 | 1855 |

[a] Reference 22. [b] Reference 21. [c] Reference 11. [d] Recent experimental data taken from reference 28 yield $D_0 = 6.305$ eV, to which one must add a ZPE of 0.115 eV, taken from http://www.cccbdb.nist.gov.

***Table 3.*** The Weights of VB Structures Gathered by Groups from VBSCF/6-31G* Calculations at the Equilibrium C−C Distance and at 2.0 Å

| structures | eq 2 ($R_{eq}$) | eq 3 ($R_{eq}$) | eq 4 ($R_{eq}$) | eq 3 (2.0 Å) | eq 4 (2.0 Å) |
|---|---|---|---|---|---|
| group 1 | 0.628 | 0.472 | 0.679 | 0.203 | 0.182 |
| group 2 | −0.052 | 0.126 | 0.017 | 0.199 | 0.206 |
| group 3 | 0.264 | 0.248 | 0.200 | 0.173 | 0.095 |
| group 4 | 0.061 | 0.048 | 0.051 | 0.111 | 0.314 |
| group 5 | 0.080 | 0.074 | 0.043 | 0.073 | 0.037 |
| group 6 | 0.018 | 0.033 | 0.011 | 0.070 | 0.040 |
| 6-$\sigma$-2$\pi$ | 0.000 | 0.001 | 0.000 | 0.172 | 0.127 |

value. The VBCIS-computed results bring significant improvements relative to the VBSCF method. The VBCIS/6-31G* computed dissociation energy is the value of 6.29 eV, slightly overestimating the bonding energy with respect to full CI in the same basis set. This is a systematic tendency of VBCIS and related methods that are generally found to yield bonding energies intermediate between full CI and experimental values, owing to a slight excess of relative correlation energy that "fortunately" compensates for the paucity of the basis set.[13] The most accurate VB computed vibrational frequency in Table 1, 1895 $cm^{-1}$ in the cc-pVTZ basis set, is within 2% of the experimental data. All of these successful VB results show that the VB description of the $C_2$ ground state is accurate, despite the small number of VB structures, and that the subsequent analysis of the bonding mode can be trusted.

**Nature of the $X^1\Sigma_g^+$ Ground State.** Table 3 shows the VBSCF/6-31G* weights of the various VB structures, summed up for each group, as calculated by eqs 2−4. As written already, the various formulas (eqs 2−4) lead to different values but very similar trends and have therefore a qualitative significance.

In all of the weight definitions (eqs 2−4), group 1, which represents the triply bonded structure, is the major bonding mode at the equilibrium distance with weights of 0.472−0.679 (Table 3, columns 2−4). At the same geometry, the second group, by order of importance, is group 3, which describes a $\sigma + \pi$ double bond (weights 0.200−0.264). The weights for group 2, with the "suspended double $\pi$-bonding", are small and more variable, −0.05, 0.13, and 0.02. As has been mentioned in the Theory and Methodology section, negative weights are unphysical but may be found with the Chirgwin−Coulson definition of the weights (eq 2), in which they signify the low importance of the VB structure in question in the total wave function. This is confirmed by the low weights also found for the VB structures of group 2, as calculated by the alternative definitions (eqs 3 and 4). Last, group 7, the 6-$\sigma$−2-$\pi$ bonding mode, is found to be totally negligible at the equilibrium distance by all weight definitions, in harmony with its negligible stabilizing effect (vide supra) and with the fact that the VB structures belonging to this group are strictly orthogonal to all of the others. As for groups 4−6, which are necessary to ensure correct dissociation but do not correspond to chemically significant bonding schemes, their weights are found to be small, albeit not negligible, at the equilibrium geometry. Thus, if we restrict the discussion to groups 1−3 and 7, all weight definitions ($W$) end up with a clear qualitative ordering of importance in the $C_2$ ground state at the equilibrium distance, eq 6:

$$W(\text{group 1}) > W(\text{group 3}) > W(\text{group 2}) \gg W(\text{group 7}) \sim 0.0 \quad (6)$$

The description of the $C_2$ ground state in terms of a major triply bonded structure **1** and a less important, but significant, $\sigma + \pi$ doubly bonded structure **3** nicely accounts for the bond length of 1.240 Å, intermediate between that of ethylene (1.339 Å) and that of acetylene (1.203 Å), and closer to the latter.[2]

The ordering displayed in eq 6 is completely reshuffled at 2.0 Å, a distance at which the 6-$\sigma$−2-$\pi$ bonding mode (group **7**) is found to be important with all definitions (Table 3, columns 5 and 6), in agreement with high level MO studies which stress the importance of the configuration $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^2 3\sigma_g^2\rangle$ in this region of the PEC.

Another way to estimate the importance of the different groups for the description of the ground state is to calculate their individual energies at the equilibrium geometry. Even if there is no strict relationship between the energy of a VB structure and its weight in the ground state, it is generally assumed that the lower the energy of a VB structure or group-structure, the stronger will be its contribution to the ground state. The calculation of the energies of each group was performed by separate VB-CI involving only the VB structures of this group, while keeping unchanged the orbitals of the ground state. The results, as calculated at the equilibrium geometry, are displayed in Table 4, columns 2 and 3. It can be seen that group 1 is by far the lowest, followed by group 3 at 81.36 kcal/mol higher. Group 2 lies even higher, 36.82 kcal/mol over group 3. Finally, groups 4−6 are much higher, confirming their low contribution to the ground state in this geometry. Thus, it can be seen that the energy ordering of groups 1−3, in eq 7 below, is intuitively in agreement with the weight ordering in eq 6.

$$E(\text{group 1}) < E(\text{group 3}) < E(\text{group 2}) \quad (7)$$

Still another way to assess the relative contributions of each group to the description of the ground state is to calculate the stabilization energy that is brought about when this group is mixed with the major group 1. The results, obtained by separate VB-CI at the equilibrium geometry as

***Table 4.*** Energies of the Individual Groups and Their Combinations with Group 1[a]

| group | *E* (Hartrees) | *E* (kcal/mol) | groups | *E* (Hartrees) | *E* (kcal/mol) |
|---|---|---|---|---|---|
| group 1 | −75.568055 | 0 | group 1 | −75.568055 | 0 |
| group 2 | −75.379708 | 118.18 | groups 1 + 2 | −75.571034 | −1.88 |
| group 3 | −75.438395 | 81.36 | groups 1 + 3 | −75.591799 | −14.91 |
| group 4 | −75.201470 | 230.0 | groups 1 + 4 | −75.578180 | −6.36 |
| group 7 | −75.165757 | 252.44 | groups 1 + 5 | −75.572812 | −3.00 |
| group 6 | −74.941845 | 392.94 | groups 1 + 6 | −75.574775 | −4.23 |

[a] The orbitals of the ground state are used in all VB structures.

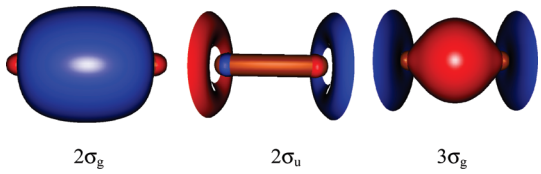



***Figure 2.*** Shapes of the $2\sigma_g$, $2\sigma_u$, and $3\sigma_g$ molecular orbitals at the equilibrium geometry from a TCSCF calculation involving the $2\sigma_g^2 2\sigma_u^2 1\pi_x^2 1\pi_y^2$ and $2\sigma_g^2 1\pi_x^2 1\pi_y^2 3\sigma_g^2$ configurations.

above, are displayed in Table 4, columns 5 and 6. It is clear that it is group 3 that contributes the most to stabilizing the ground state by mixing with group 1, with a stabilization energy of ca. 15 kcal/mol, vs only 2−6 kcal/mol for each of the other groups. On the other hand, group 2 contributes very weakly to the stabilization energy of the ground state relative to group 1, by barely 2 kcal/mol. It is therefore clear that the three types of test calculations that have been done are consistent with each other and that the emerging bonding picture of $C_2$ involves the major triply bonded structure **1**, with an important contribution of the $\sigma + \pi$ doubly bonded structure **3**, and a minor contribution of the $\pi + \pi$ structure **2**.

How can the unimportance of group 2 at the equilibrium geometry be reconciled with the MO diagram in Scheme 1, displaying a pair of occupied MOs of the $\sigma$ type, one bonding ($2\sigma_g$) and one possibly antibonding ($2\sigma_u$)? As argued in the Introduction, this can be interpreted by the shapes of the respective orbitals and by the natural orbital occupation numbers obtained from two-configuration MCSCF or CASSCF(8,8) wave functions. If the $2\sigma_g$ and $2\sigma_u$ MOs are clearly bonding and antibonding, respectively, then this argues in favor of the absence of $\sigma$ bonding, and therefore in support of the bonding mode displaying two "suspended" $\pi$ bonds, i.e., group 2. If, on the other hand, the $2\sigma_u$ orbital is only weakly antibonding, then structures displaying $\sigma$ bonding, i.e., group 1 and group 3, are favored.

Figure 2 displays the shapes of the $2\sigma_g$, $2\sigma_u$, and $3\sigma_g$ MOs, as they arise from a two-configuration MCSCF calculation involving the two most important configurations, $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^4\rangle$ and $|(\text{core})2\sigma_g^2 1\pi_u^4 3\sigma_g^2\rangle$. It appears very clearly that $2\sigma_g$ is strongly bonding, as is $3\sigma_g$, albeit to a somewhat lesser extent than $2\sigma_g$. On the other hand, the $2\sigma_u$ MO is an out-of-phase combination of two outward-directed hybrids and has therefore very little antibonding character. It follows that the two most important configurations in the $X^1\Sigma_g^+$ ground state both display some important contribution from $\sigma$ bonding, in agreement with the predominant weights of group 1 and group 3 structures arising from the VB calculations, and ruling out group 2.

**Diabatic Potential Energy Curves.** Figure 3 shows the diabatic potential energy curves (PECs) for each individual

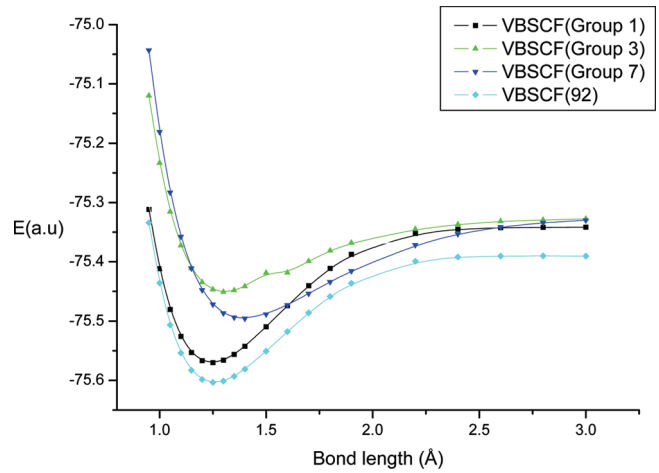



***Figure 3.*** The VBSCF calculated PECs for the ground states of the $C_2$ molecule with the 6-31G* basis set.

group of VB structures, shown above in Scheme 3, calculated at the VBSCF/6-31G* level. For each group, the VBSCF calculation is performed in the space of all of the VB structures that belong to this group, and this time the orbitals are reoptimized for each group, so as to get the lowest possible energies for the diabatic curves (VDC method, see Theory and Methodology section). Groups 4−6 that are included in the calculation of the ground state for completeness but do not represent chemically meaningful bonding modes are not considered in Figure 3. Moreover, we encountered some computational difficulties with the diabatic energy curve of group 2, which was found to lie between group 1 and group 3 at the equilibrium distance, in contradiction with the results displayed in Table 4 (group 2 lying above group 3). However, examination of the VB wave functions showed that some of the orbitals optimized for group 2 alone are exceedingly different from the corresponding orbitals in the ground state. For example, the overlap between orbitals $p_{1z}$ and $p_{2z}$ (Scheme 2) is 0.70 in the ground state, vs only 0.30 as optimized for group 2 alone. It was therefore concluded that the result of the VBSCF calculation of group 2 alone has little to do with group 2 as a component of the ground state. For this reason, Figure 3 is restricted to the diabatic curves of groups 1, 3, and 7, together with the adiabatic ground state.

It is seen that the lowest diabatic curve in the region near the equilibrium geometry is that of group 1, which represents the triply bonded bonding mode. Group 3 and group 7 curves lie significantly higher than those of group 1 at the equilibrium geometry, but group 7 (the 6-$\sigma$−2-$\pi$ bonding mode) becomes the lowest one at 1.7 Å, in agreement with the full CI study in the MO framework that predicts that determinants

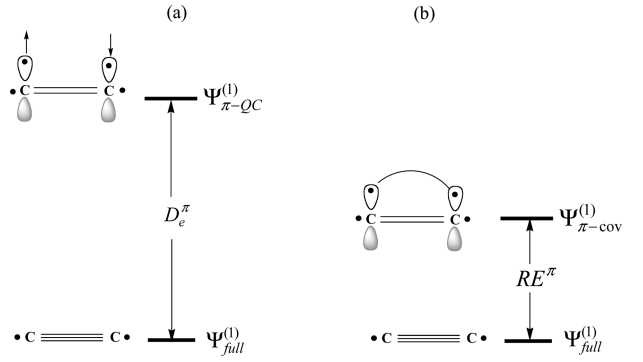



***Figure 4.*** (a) Definition of in situ bond energy for a $\pi$ bond of structure **1** based on the quasiclassical state (QC) reference. (b) Definition of the covalent-ionic resonance energy for a $\pi$ bond of structure **1**.

of the type $|(\text{core})2\sigma_g^2 2\sigma_u^2 1\pi_u^2 3\sigma_g^2\rangle$ become dominant from this distance onward.

**Characteristics of the Individual $\sigma$ and $\pi$ Bonds.** As has been shown above, the VBCIS calculation with a restricted set of 78 structures yields an accurate total bond dissociation energy (BDE), calculated as a difference between the molecular energy and the sum of atomic energies. However, this global quantity does not give us much information on the individual $\sigma$ and $\pi$ bonds that participate in the overall bonding. Two questions of interest are as follows: (i) What are the relative strengths of the $\pi$ and $\sigma$ bonds as contributors to the overall bonding? (ii) What are the natures of these individual bonds: classical covalent or, rather, charge-shift bonds (see the definition below)? These features of the $\sigma$ and $\pi$ bonds were investigated by restricting attention to the major VB structure (**1**).

In multiple bonds, like in $C_2$, it is not simple to experimentally determine separate bond energies for $\sigma$ and $\pi$ bonds. In double bonds (e.g., ethylene), one can roughly estimate the $\pi$ bonding energy as the rotational barrier; however, this quantity also involves relaxation of the $\sigma$ bond and hyperconjugation in the twisted form. The problem gets even worse with triple bonds, since these species do not have rotational barriers. However, as we have shown amply before,[13,40−43] these difficulties can be bypassed by defining *a nonbonded reference state* for one $\pi$ bond, in which the two electrons maintain opposite spins but do not exchange. For example, if one wants to estimate the strength of one of the two $\pi$ bonds in the triply bonded structure **1**, one may define a nonbonded state, called the "quasiclassical state" $\Psi^{(1)}_{\pi\text{-QC}}$ in Figure 4, in which the electrons of a given $\pi$ bond have only one spin arrangement pattern (only $\alpha\beta$), and therefore this structure by itself has no bonding due to the resonance with the second spin arrangement pattern that is required to form a singlet pair. In such a state, the interactions across the unpaired $\pi$ bond in $\Psi^{(1)}_{\pi\text{-QC}}$ involve only classical electron−electron repulsion, nuclear repulsion, and electron−nuclear attraction, and since the fragments are neutral, these terms sum to approximately zero. Thus, the difference between the energy of $\Psi^{(1)}_{\pi\text{-QC}}$ and that of the fully bonded structure **1** gives the "in situ" $\pi$-bonding energy $D_e^{\pi}$ at a given interatomic distance, i.e., the bonding energy that effectively stabilizes the electronic interaction of the considered $\pi$ bond without any other relaxation or reorganization term arising from other bonds. The in situ $\pi$-bonding energy in structure **1** is expressed in eq 9:

$$D_e^{\pi} = E(\Psi^{(1)}_{\pi-\text{QC}}) - E(\Psi^{(1)}_{\text{full}}) \tag{9}$$

where $\Psi^{(1)}_{\text{full}}$ is the fully optimized wave function for structure **1**, calculated by a VBSCF calculation involving the 21 VB structures of group 1.

An important point to note is that the QC state remains nonbonding only at distances equal to or longer than the optimal bonding distance[12] but becomes repulsive at shorter distances and therefore ceases to be a good reference state for measuring the in situ bonding energy in such a case. Moreover, one expects the repulsive wall of the QC state to be basis set dependent: the more flexible the basis set, the less repulsive the QC state. Accordingly, the method can be applied for the $\pi$ components of multiple bonds but will be much less accurate for the $\sigma$ component, since at a length of 1.25 Å, the latter is "compressed" relative to the optimal distance for a single $\sigma$ bond ($\sim$1.50 Å for a C−C $\sigma$ bond between sp hybrids). Thus, we will only get a rough estimation for the $\sigma$ bond strength.

Equation 9 has been applied for structure **1** with the 6-31G* basis set, yielding an in situ bonding energy of 93.3 kcal/mol for the $\pi$ bond, very close to the value 92.25 kcal/mol that was found with the same technique by Ploshnik[44] for the $\pi$ bond of acetylene. Interestingly, these values are significantly larger than the in situ $\pi$ bonding energy of 72.0 kcal/mol calculated by Galbraith et al.[6] for ethylene, in agreement with the principle that $\pi$ bonds prefer short bond lengths.

The same technique has been used to estimate the in situ bonding energy of the $\sigma$ bond in **1**, yielding an in situ bonding energy of 99.4 kcal/mol in 6-31G* basis set, and only 64.1 kcal/mol in the larger cc-pVTZ one, thus confirming the expected basis set dependency and the lack of accuracy of the "in situ" estimation of the $\sigma$ bond strength of $C_2$.

Summing up the $\sigma$ and $\pi$ in situ bonding energies, one would arrive at a total of 251−286 kcal/mol for structure **1**, a value that is of course much larger than the true dissociation energy because the fragments enjoy some demotion energy from their local high spin states in the molecule to their triplet states at infinite distance. This high value can be compared to the estimated bond strength of acetylene relative to the high spin fragments, as calculated by Frenking.[45] In this approach, which is also that of Trinquier and Malrieu,[46,47] Carter and Goddard,[48] and others,[49] the acetylene molecule is considered as the product of interactions between two $^4\Sigma^-$ CH fragments, yielding a triple bonding energy of 270.9 kcal/mol,[45] before the final dissociation energy is obtained after adding the demotion energy of the fragments from $^4\Sigma^-$ to their ground states $^2\Pi$. From this comparison between acetylene and the triply bonded structure (**1**) of $C_2$, it seems that our window of 64−99 kcal/mol for the in situ $\sigma$ bond strength of structure **1** is reasonable.

In the VB framework, any two-electron bond is described as a superposition of one covalent and two ionic forms, even

***Table 5.*** In Situ $\pi$-Bonding Energies and Covalent-Ionic Resonance Energies for the $\pi$ and $\sigma$ Bonds of the Triply Bonded Structure **1**, As Calculated by the VBSCF Method (Energies in kcal/mol)

| basis set | $D_e^{\pi}$ | $D_e^{\sigma}$ | $RE^{\pi}$ | $RE^{\sigma}$ |
|---|---|---|---|---|
| 6-31G* | 93.3 | 99.4 | 44.2 | 15.1 |
| cc-pVTZ | 88.1 | 64.1 | 36.1 | 12.2 |

in the homonuclear case.[13] Generally speaking, whenever the covalent structure has the largest weight, which is always the case for homonuclear molecules,[50] the bond dissociation energy BDE of any of the $\sigma$ or $\pi$ bonds in structure **1** will be given in eq 10:

$$\mathrm{BDE} = \mathrm{BDE}_{\mathrm{cov}} + \mathrm{RE}_{\mathrm{cov\text{-}ion}} \tag{10}$$

Here, $BDE_{cov}$ is the covalent spin-pairing energy of the bond, and $RE_{cov\text{-}ion}$ is the covalent-ionic resonance energy due to the mixing of the ionic structures into the covalent one(s). Both quantities are variational within the subset of VB structures. In classical covalent bonds (e.g., $H_2$, $H_3C-CH_3$, etc.), the covalent term is the major one; however, there exists a category of bonds, termed "charge-shift" (CS) bonds, where it is the resonance energy term, $RE_{cov\text{-}ion}$, that is the major one, in eq 10, and responsible for most of the bonding.[50,51]

The covalent-ionic resonance energy of a $\pi$ bond in structure **1** can be estimated by means of eq 11, as schematized in Figure 4b:

$$\mathrm{RE}^{\pi} = E(\Psi^{(1)}_{\pi-\mathrm{cov}}) - E(\Psi^{(1)}_{\mathrm{full}}) \tag{11}$$

where $\Psi^{(1)}_{\pi\text{-cov}}$ is a variational combination of all of the VB structures in group 1 in which the considered $\pi$ bond is purely covalent, and $\Psi^{(1)}_{\mathrm{full}}$ is defined as in eq 9. The calculated values are displayed in Table 5. The 6-31G* value for $RE^{\pi}$, 44.2 kcal/mol per $\pi$ bond, is seen to be slightly smaller than 50% of the in situ $\pi$-bonding energy (93.3 kcal/mol), showing that the $\pi$ bonds in **1** have a strong charge-shift character and lie between classical covalent bonds and charge-shift ones, according to our classification.[50] The same conclusion is reached with the calculation in the cc-pVTZ basis set.

The charge-shift character of the $\sigma$ bond in structure **1** can be estimated in an analogous way, by means of eq 12:

$$\mathrm{RE}^{\sigma} = E(\Psi^{(1)}_{\sigma-\mathrm{cov}}) - E(\Psi^{(1)}_{\mathrm{full}}) \tag{12}$$

where $\Psi^{(1)}_{\sigma\text{-cov}}$ now involves the VB structures in group 1 in which the $\sigma$ bond is purely covalent. The $RE^{\sigma}$ values, 12−15 kcal/mol (Table 5), are now much smaller than the estimated in situ $\sigma$-bonding energy, clearly classifying the $\sigma$ bond in **1** as a classical covalent bond.

## Conclusion

The electronic structure of the $C_2$ molecule in the ground state was described herein using the ab initio VB calculations. While this molecule is known to require very sophisticated computational methods in the MO−CI framework, owing to its strong multireference character, the VB method encounters no particular difficulties with this challenging molecule. Thus, a simple VBSCF calculation involving 78 VB structures, selected on the basis of chemical criteria, is followed by configuration interaction with single excitations (VBCIS). Such a calculation provides spectroscopic parameters that are very close to experimental values in the largest basis set, and also close to full CI calculated values in the smallest one.

According to the VB results, the electronic structure of the $C_2$ ground state is more complex than that of acetylene and is described in terms of three interacting bonding schemes. A triply bonded structure, analogous to the bonding scheme of acetylene, is the major one in terms of weights, and the lowest in energy at the equilibrium distance. The second structure in terms of weights is an ethylene-like structure, displaying a $\sigma + \pi$ double bond. The structure with two suspended $\pi$ bonds but no $\sigma$ bond contributes only marginally to the ground state.

The natures of the $\pi$ and $\sigma$ bonds are investigated in the triply bonded bonding scheme. The "in situ" strength of each of these bonds is estimated, defined as the effective strength of the bonding interaction at the equilibrium distance without the geometrical relaxation and fragment reorganization that occur at large distances. The $\pi$ bond is found to be stronger than the $\pi$ bond of ethylene, but as strong as the $\pi$ bond of acetylene with an in situ bond strength of 93.3 kcal/mol. The $\sigma$ bond strength could not be estimated accurately by the "in situ" technique; however the estimated order of magnitude shows that this bond strongly contributes to the overall bonding in $C_2$. This probably explains why the structure with suspended $\pi$ bonds cannot be the major one in the ground state of $C_2$.

The VB calculations also allow for specification of the contribution of the covalent-ionic resonance energy to bonding in each of the $\pi$ and $\sigma$ bonds. While the $\sigma$ bond is found to be a perfectly classical covalent bond, the $\pi$ bonds have significant contribution from the covalent-ionic resonance energy, which classifies them as intermediate between classical covalent and charge-shift bonds.

Finally, the bonding picture of $C_2$ that emerges from this study is that of a mixture of a major triply bonded structure, •C≡C•, with rather strong $\sigma$ and $\pi$ bonds, perturbed by a less important $\sigma + \pi$ doubly bonded structure :C═C:, which accounts for the bond length being intermediate between that of ethylene and that of acetylene, but closer to the latter molecule.

**Acknowledgment.** This project is supported by the Natural Science Foundation of China (Nos. 20873106, 21003101). S.S. is supported by an Israel Science Foundation Grant, ISF 53/09. This paper is dedicated to E.D. Jemmis on occasion of his forthcoming 60th birthday.

**Note Added in Proof.** We thank one of the referees for pointing out the results of some unpublished CASSCF (8,8) calculations in support of the triply bonded structure **1**. The occupation numbers of the natural orbitals from this calculation indicate that although the $\pi$ bonds (x and y) are very important, with occupation numbers of 1.89 for the bonding orbitals vs 0.12 for the antibonding ones, the $\sigma$ orbitals also contribute to bonding, with cumulated occupa-

tion numbers of 2.38 for the bonding $\sigma_g$ orbitals vs only 1.61 for the antibonding $2\sigma_u$ which, we recall, is only slightly antibonding

**Supporting Information Available:** These five schemes show all 92 VB structures in groups 1−7. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References


(1) Jemmis, E. D.; Pathak, B.; King, R. B.; Schaefer, H. F., III. *Chem. Commun.* **2006**, *20*, 2164.

(2) NIST Standard Reference Database 101. CCCBDB on the web. http://cccbdb.nist.gov (accessed April 2010).

(3) Van Nes, G. J. H.; Vos, A. *Acta Crystallogr., Sect. B* **1978**, *34*, 1947.

(4) Van Nes, G. J. H.; Vos, A. *Acta Crystallogr., Sect. B* **1979**, *35*, 2593.

(5) McMullan, R. K.; Kvick, Å.; Popelier, P. *Acta Crystallogr., Sect. B* **1992**, *48*, 726.

(6) Galbraith, J. M.; Blank, E.; Shaik, S.; Hiberty, P. C. *Chem.—Eur. J.* **2000**, *6*, 2425.

(7) Pyykkö, P.; Riedel, S.; Patzschke, M. *Chem.—Eur. J.* **2005**, *11*, 3511.

(8) Huntley, D. R.; Markopoulos, G.; Donovan, P. M.; Scott, L. T.; Hoffmann, R. *Angew. Chem., Int. Ed.* **2005**, *44*, 7549.

(9) Abrams, M. L.; Sherrill, C. D. *J. Chem. Phys.* **2003**, *118*, 1604.

(10) Abrams, M. L.; Sherrill, C. D. *J. Chem. Phys.* **2004**, *121*, 9211.

(11) Sherrill, C. D.; Piecuch, P. *J. Chem. Phys.* **2005**, *122*, 124104.

(12) Kutzelnigg, W. In *Theoretical Models of Chemical Bonding*; Maksic, Z. B., Ed.; Springer-Verlag: Berlin, Germany, 1990; Vol. 2, pp 1−43.

(13) Shaik, S.; Hiberty, P. C. *A Chemist's Guide to Valence Bond Theory*; Wiley-Interscience: New York, 2007.

(14) Douay, M.; Nietmann, R.; Bernath, P. F. *J. Mol. Spectrosc.* **1988**, *131*, 261.

(15) Douay, M.; Nietmann, R.; Bernath, P. F. *J. Mol. Spectrosc.* **1988**, *131*, 250.

(16) Boggio-Pasqua, M.; Halvick, P.; Rayez, M. T.; Rayez, J. C.; Robbe, J. M. *J. Phys. Chem. A* **1998**, *102*, 2009.

(17) Weltner, W.; Van Zee, R. J. *Chem. Rev.* **1989**, *89*, 1713.

(18) Martin, M. *J. Photochem. Photobiol. A* **1992**, *66*, 263.

(19) Van Orden, A.; Saykally, R. J. *Chem. Rev.* **1998**, *98*, 2313.

(20) Boggio-Pasqua, M.; Voronin, A. I.; Halvick, P.; Rayez, J. C. *THEOCHEM* **2000**, *531*, 159.

(21) Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1992**, *96*, 6073.

(22) Pradhan, A. D.; Partridge, H.; Bauschlicher, J. C. W. *J. Chem. Phys.* **1994**, *101*, 3857.

(23) Peterson, K. A. *J. Chem. Phys.* **1995**, *102*, 262.

(24) Varandas, A. J. C.; da Silva, j. D. *J. Chem. Soc., Faraday Trans. 2* **1986**, *82*, 593.

(25) Varandas, A. J. C.; da Silva, J. D. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 941.

(26) Varandas, A. J. C. *J. Chem. Phys.* **2008**, *129*, 234103.

(27) Mahapatra, U. S.; Chattopadhyay, S.; Chaudhuri, R. K. *J. Chem. Phys.* **2008**, *129*, 024108.

(28) Bytautas, L.; Ruedenberg, K. *J. Chem. Phys.* **2005**, *122*, 154110.

(29) Wu, W.; Mo, Y.; Cao, Z.; Zhang, Q. In *Valence Bond Theory*; Cooper, D. L., Ed.; Elsevier: Amsterdam, The Netherlands, 2002; pp 143−185.

(30) van Lenthe, J. H.; Balint-Kurti, G. G. *J. Chem. Phys.* **1983**, *78*, 5699.

(31) van Lenthe, J. H.; Balint-Kurti, G. G. *Chem. Phys. Lett.* **1980**, *76*, 138.

(32) Song, L.; Wu, W.; Zhang, Q.; Shaik, S. *J. Comput. Chem.* **2004**, *25*, 472.

(33) Wu, W.; Song, L.; Cao, Z.; Zhang, Q.; Shaik, S. *J. Phys. Chem. A* **2002**, *106*, 2721.

(34) Chirgwin, B. H.; Coulson, C. A. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196.

(35) Löwdin, P.-O. *Ark. Mat. Astr. Fys.* **1947**, *A35*, 9.

(36) Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *J. Comput. Chem.* **2005**, *26*, 514.

(37) Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *XMVB*, version 1.0; Xiamen University: Xiamen, China, 2003.

(38) Frisch, M. J. *Gaussian 03*, Revision D.01; Gaussian: Wallingford, CT, 2004.

(39) Weyl, H. In *The Theory of Groups and Quantum Mechanics*; Dover Publications: New York, 1956.

(40) Hiberty, P. C.; Danovich, D.; Shurki, A.; Shaik, S. *J. Am. Chem. Soc.* **1995**, *117*, 7760.

(41) Jug, K.; Hiberty, P. C.; Shaik, S. *Chem. Rev.* **2001**, *101*, 1477.

(42) Wu, W.; Gu, J.; Song, J.; Shaik, S.; Hiberty, P. *Angew. Chem., Int. Ed.* **2009**, *48*, 1407.

(43) Shaik, S.; Chen, Z.; Wu, W.; Stanger, A.; Danovich, D.; Hiberty, P. C. *ChemPhysChem* **2009**, *10*, 2658.

(44) Ploshnik, E. MSc Thesis, The Hebrew University, Jerusalem, Israel, 2005.

(45) Lein, M.; Krapp, A.; Frenking, G. *J. Am. Chem. Soc.* **2005**, *127*, 6290.

(46) Trinquier, G.; Malrieu, J. P. *J. Am. Chem. Soc.* **1987**, *109*, 5303.

(47) Malrieu, J. P.; Trinquier, G. *J. Am. Chem. Soc.* **1989**, *111*, 5916.

(48) Carter, E. A.; Goddard, W. A. *J. Phys. Chem.* **1986**, *90*, 998.

(49) Sugiyama, Y.; Sasamori, T.; Hosoi, Y.; Furukawa, Y.; Takagi, N.; Nagase, S.; Tokitoh, N. *J. Am. Chem. Soc.* **2005**, *128*, 1023.

(50) Shaik, S.; Danovich, D.; Silvi, B.; Lauvergnat, D. L.; Hiberty, P. C. *Chem.—Eur. J.* **2005**, *11*, 6358.

(51) Shaik, S.; Danovich, D.; Wu, W.; Hiberty, P. C. *Nat. Chem.* **2009**, *1*, 443.


CT100577V

JCTC Journal of Chemical Theory and Computation

# Substantial Dissociation Energies for the Recently Synthesized NC−Ag−$NH_3$ and Br−Ag−$NH_3$ Molecules and Their Isovalent Family Members M(CN)$XY_3$ and M(Br)$XY_3$ (M = Cu, Ag, Au; X = N, P; Y = H, F)

Qiong Luo,*[,†,‡] Qianshu Li,[†,‡] Yaoming Xie,[§] R. Bruce King,[‡,§] and Henry F. Schaefer*[,§]

*State Key Laboratory of Explosion Science and Technology, Beijing Institute of Technology, Beijing, 100081, P. R. China*

*Center for Computational Quantum Chemistry, South China Normal University, Guangzhou, 510631 China*

*Department of Chemistry and Center for Computational Quantum Chemistry, University of Georgia, Athens, Georgia 30602, United States*

Received October 10, 2010

**Abstract:** Chippindale et al. have recently synthesized the unique molecules (NC)Ag($NH_3$) and BrAg($NH_3$) and shown the heavy atom skeletal structures to be linear. Here, a theoretical study is reported of 12 members each of the two isovalent series of molecules. For (NC)Ag($NH_3$) and BrAg($NH_3$), the theoretical structures agree well with those determined by X-ray crystallography. Structures for the 22 yet unknown compounds should be similarly reliable. The dissociation energies for a loss of $NH_3$ from the two known compounds are significant (34 and 31 kcal/mol), confirming their viability. For the other systems, the ligand dissociation energies are highly variable, ranging from 9 kcal/mol (BrAg-$NF_3$) to 44 kcal/mol (BrAu-$PH_3$). The bond dissociation energies for the different metals follow the irregular order Au > Cu > Ag. For the $XY_3$ ligands, the dissociation energies follow the order $NH_3$ > $PH_3$ > $PF_3$ > $NF_3$, except for the BrAu−$XY_3$ complexes. Electronic structure insights are gained via Natural Bond Orbital (NBO) analyses.

## Introduction

The metals most frequently forming linear two-coordinate metal complexes are the $d^{10}$ metals, particularly the coinage metals Cu, Ag, and Au in their +1 oxidation states.[1] For gold, this linear coordination chemistry dominates the +1 oxidation state, and both symmetrical and unsymmetrical linear Au(I) complexes are known. For example, phosphine complexes of gold(I) halides, namely, $R_3PAuX$ (X = Cl, Br, I), are commonly used reagents in gold chemistry.[2,3] Such linear gold(I) complexes are relatively inert kinetically. They are rather unreactive toward increasing their coordination number above two by ligand addition reactions. This characteristic feature of linear Au(I) chemistry may be attributed to relativistic effects.[2]

The situation is different with the lighter coinage metals copper and silver. Although the organic bis(silver carbene) complexes, such as Ag-based N-heterocyclic carbene complexes, have been reported,[4−12] before 2008, simple linear asymmetrical complexes of Cu(I) and Ag(I) were unknown because of limitations in synthetic methods, the kinetic lability of such complexes, and the formation of infinite chain complexes or species with higher metal coordination numbers.[13−19] However, symmetrical linear two-coordinate complexes of copper and silver have been known for more than a century. Using silver as an example, the symmetrical

* Corresponding authors. E-mail: kellyluo@bit.edu.cn (Q.L.); sch@uga.edu (H.F.S.).

† Beijing Institute of Technology.

‡ South China Normal University.

§ University of Georgia.

10.1021/ct100664q 
Published on Web 12/09/2010

linear complexes $Ag(NH_3)_2^+$ and $AgX_2^-$ (X = CN, Cl, Br, I) have been known for many decades and are routinely discussed in inorganic chemistry textbooks.[1]

A recent (2008) study of the behavior of silver(I) derivatives in aqueous ammonia reported a breakthrough in the synthesis of asymmetrical linear silver(I) derivatives.[13] Therein, cautious evaporation of solvent from concentrated solutions of silver(I) halides or pseudohalides in aqueous ammonia led to the derivatives NC−Ag−$NH_3$ and Br−Ag−$NH_3$, which were shown by X-ray crystallography to be discrete molecules rather than infinite chains. The cyanide complex NC−Ag−$NH_3$ was found to be stable to 100 °C before losing ammonia to form AgCN. The bromide complex Br−Ag−$NH_3$ was less stable, losing ammonia around room temperature to give AgBr. It thus becomes of interest to investigate factors affecting the stability of NC−Ag−$NH_3$ and other complexes analogous to it. Recently, the gas-phase monoammoniate of silver chloride ($NH_3$···Ag−Cl) was also observed by Mikhailov et al.,[20] and the corresponding *ab initio* calculations performed by the authors.

This paper describes our attempts using density functional theory to explore the possible range of coinage metal derivatives of the type NC−M−$XY_3$ and Br−M−$XY_3$ (M = Cu, Ag, Au; X = N, P; Y = H, F).

## Theoretical Methods

Two density functional theory (DFT) or hybrid Hatree-Fock/DFT methods were used in this study. The first functional is B3LYP, which incorporates Becke's three-parameter functional (B3)[21] with the Lee, Yang, and Parr (LYP) correlation functional.[22] The second approach is BP86, using the exchange functional of Becke[23] in conjunction with the correlation functional of Perdew.[24]

For the coinage metals, Cu, Ag, and Au, we adopted the Stuttgart/Dresden double-ζ (SDD) effective core potential (ECP) basis sets.[25] In these basis sets, the innermost-electrons (10 for Cu, 28 for Ag, and 60 for Au) for the transition metal atoms are replaced by the effective core potentials (ECP), which include relativistic effects, known to be important for the heavy transition metal atoms.[2] For the three coinage metals, the SDD basis sets for the 19 valence electrons are Cu(8s7p6d/6s5p3d), Ag(8s7p6d/6s5p3d), and Au(8s6p5d/7s3p4d).

All-electron double-ζ plus polarization basis sets augmented with diffuse functions (DZP++) are used for main group elements. The DZP++ basis sets for carbon, nitrogen, fluorine, and phosphorus are composed of the standard Huzinaga−Dunning double-ζ basis sets[25−28] plus one set of pure spherical harmonic d functions with orbital exponents $\alpha_d(C) = 0.75$, $\alpha_d(N) = 0.80$, $\alpha_d(F) = 1.00$, and $\alpha_d(P) = 0.60$ augmented with one set of diffuse functions $\alpha_s(C) = 0.04302$ and $\alpha_p(C) = 0.03629$, $\alpha_s(N) = 0.06029$, and $\alpha_p(N) = 0.05148$, $\alpha_s(F) = 0.10490$ and $\alpha_p(F) = 0.08260$, and $\alpha_s(P) = 0.03448$ and $\alpha_p(P) = 0.03346$. For H, the added polarization functions are one set of p-type functions with orbital exponent $\alpha_p(H) = 0.75$, augmented with one diffuse s function $\alpha_s(H) = 0.04415$. For bromine, the basis set was composed of Ahlrichs' standard double-spd set plus a set of d-type polarization functions $\alpha_d(Br) = 0.389$ plus diffuse functions $\alpha_s(Br) = 0.0469$ and $\alpha_p(Br) = 0.0465$.[29] The final contracted basis sets are thus designated as H(5s1p/3s1p), [C,N,F](10s6p1d/5s3p1d), P(12s8p1d/7s5p1d), and Br(15s12p6d/9s7p3d).

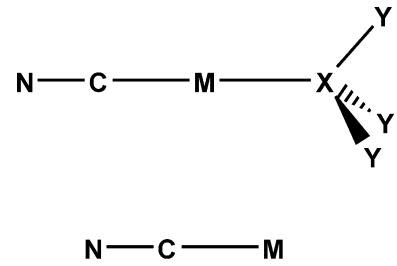



***Figure 1.*** Qualitative structures of M(CN)($XY_3$) and MCN (M = Cu, Ag, Au; X = N, P; Y = H, F).

***Table 1.*** Geometrical Parameters for M(CN)$NH_3$ (M = Cu, Ag, Au)[a]

| | | B3LYP | | | BP86 | | |
|---|---|---|---|---|---|---|---|
| species | | Cu | Ag | Au | Cu | Ag | Au |
| M(CN)($NH_3$) | M−C | 1.844 | 2.021 | 1.962 | 1.825 | 1.997 | 1.948 |
| | C−N | 1.174 | 1.173 | 1.172 | 1.187 | 1.186 | 1.186 |
| | M−N | 1.954 | 2.186 | 2.142 | 1.943 | 2.162 | 2.132 |
| | N−H | 1.023 | 1.022 | 1.022 | 1.031 | 1.029 | 1.030 |
| | ∠H−N−H | 107.1 | 107.2 | 107.5 | 107.0 | 107.1 | 107.5 |

[a] Bond distances in Å, bond angles in degrees.

Natural bond orbital (NBO) analyses[30] have been carried out, with the bond orders and natural charges used to gain some understanding of the bonding in these molecules. All computation employed the Gaussian 03 program suite,[31] and all results refer to the gas phase at 0 K.

The difference between the B3LYP and BP86 results is generally small, and the two sets of predictions show the same trends. Although all results from the two methods are shown in the tables, the B3LYP results are mainly discussed in the text.

## Results and Discussion

**Ag(CN)$NH_3$.** Our theoretical structure for the isolated Ag(CN)$NH_3$ molecule is shown in Figure 1, and the corresponding optimized parameters are reported in Table 1. The equilibrium NC−Ag−$NH_3$ complex is predicted by both the B3LYP and the BP86 methods to be a linear structure with $C_{3v}$ symmetry, in agreement with the experimental crystal structure reported by Chippindale et al. in 2008.[13] Our theoretical results reproduce the experimental structure reasonably well, with the difference of bond distances less than 0.1 Å. Table 1 shows that the theoretical Ag−C distance (2.021 Å) is 0.03 Å shorter than the experimental value (2.051 Å), while the theoretical Ag−N distance (2.186 Å) is a 0.07 Å longer than the corresponding experimental conclusion (2.114 Å).

A key point for this newly observed asymmetric mononuclear Ag(I) complex Ag(CN)$NH_3$ would be the nature of the bonding between its AgCN and $NH_3$ fragments. The geometry of the $NH_3$ fragment in the Ag(CN)$NH_3$ complex is comparable with that of the free $NH_3$ molecule. Table 1 shows that the N−H distance is 1.022 Å in Ag(CN)$NH_3$,

***Table 2.*** Geometrical Parameters for Separated MCN (M = Cu, Ag, Au) and $XY_3$ (X = N, P; Y = H, F) for Comparison[a]

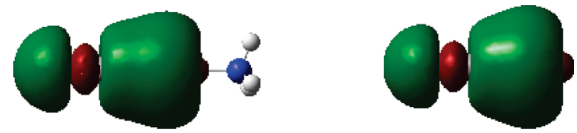
| species | | B3LYP | BP86 | expt[b] | theor[c] |
|---|---|---|---|---|---|
| $NH_3$ | N−H | 1.020 | 1.028 | | |
| | ∠H−N−H | 107.4 | 106.7 | | |
| $NF_3$ | N−F | 1.386 | 1.411 | | |
| | ∠F−N−F | 101.9 | 101.8 | | |
| $PH_3$ | P−H | 1.423 | 1.435 | | |
| | ∠H−P−H | 93.6 | 92.7 | | |
| $PF_3$ | P−F | 1.596 | 1.613 | | |
| | ∠F−P−F | 97.5 | 97.7 | | |
| CuCN | Cu−C | 1.832 | 1.807 | 1.82962(4) | 1.824 (1.8259) |
| | C−N | 1.173 | 1.187 | 1.16213(3) | 1.164 (1.1665) |
| AgCN | Ag−C | 2.037 | 2.007 | 2.031197(23) | 2.024 |
| | C−N | 1.172 | 1.186 | 1.160260(26) | 1.164 |
| AuCN | Au−C | 1.942 | 1.922 | 1.9122519(84) | 1.911 |
| | C−N | 1.171 | 1.186 | 1.1586545(97) | 1.162 |

[a] Bond distances in Å, bond angles in degrees. [b] The experimental values for CuCN refer to ref 32, and those for AgCN and AuCN refer to ref 33. [c] The theoretical values for MCN (M = Cu, Ag, and Au) were predicted at the CCSD(T)/cc-pVQZ level of theory in ref 34. The values in parentheses were predicted at the DK-CCSD(T)/cc-pVQZ level of theory (ref 35).

compared with 1.020 Å in the free $NH_3$ (Table 2). The H−N−H bond angle 107.2° in Ag(CN)$NH_3$ has only a negligible difference from that in free ammonia (107.4°). This result suggests that there is no significant orbital overlap, but rather a dipole attraction between the $NH_3$ ligand and AgCN, and this argument is supported by the NBO analysis.

The natural bond orbital (NBO) analysis provides detailed insight into these linear asymmetrical complexes of Ag(CN)-$NH_3$ (Table 3). The charge distribution of the AgCN fragment in Ag(CN)$NH_3$ bears a strong resemblance to that of the isolated AgCN molecule. The bonding of $NH_3$ to AgCN has little influence on the Wiberg Ag−C bond index, changing from 0.60 to 0.58. The highest occupied molecular orbitals (HOMO) of Ag(CN)$NH_3$ and AgCN are very similar (Figure 2). These similarities between Ag(CN)$NH_3$ and AgCN show that the Ag 5s orbital mainly takes part in the Ag−C bond, with or without the presence of $NH_3$. The Wiberg bond index of Ag−N in Ag(CN)$NH_3$ is as small as 0.20, which also suggests a small overlap between orbitals (or small covalent bonding) of the Ag and N atoms. The natural charge for the ammonia N atom is −1.12, and that for the Ag atom is +0.54 (Table 3), supporting the existence of an ionic attraction between the two fragments AgCN and $NH_3$. However, $\sigma$ donation from the ammonia lone pair to the partially vacant Ag 5s hole is not excluded.

The dissociation energy of Ag(CN)$NH_3$ to free the $NH_3$ ligand is substantial. Table 4 shows that the dissociation energy for $(NH_3)AgCN \rightarrow AgCN + NH_3$ is 33.4 (B3LYP) or 35.3 (BP86) kcal/mol. This value indicates that the simple linear asymmetric Ag(CN)$NH_3$ complex is quite favorable in energy with respect to such a dissociation. However,



Ag(CN)($NH_3$) Ag(CN)

***Figure 2.*** The HOMOs for Ag(CN)($NH_3$) and AgCN.

***Table 4.*** Reaction Energies in Kilocalories per Mole[a]

| | B3LYP | BP86 |
|---|---|---|
| $(NH_3)AgCN \rightarrow AgCN + NH_3$ | 33.4 | 35.3 |
| $(NH_3)AgCN \rightarrow 1/2[Ag(NH_3)_2]^+ + 1/2[Ag(CN)_2]^-$ | −32.1 | −32.8 |
| $2(NH_3)AgCN \rightarrow (NH_3)AgCN{-}AgCN + NH_3$ | −7.7 | −6.9 |
| $(NH_3)AgCN + AgCN \rightarrow (NH_3)AgCN{-}AgCN$ | −41.1 | −42.2 |
| $2AgCN \rightarrow AgCN{-}AgCN$ | −36.6 | −37.0 |

[a] All of the molecules are in the gas phase.

compared with the long-known symmetrical complexes $[Ag(NH_3)_2]^+$ and $[Ag(CN)_2]^-$, the asymmetric NC−Ag−$NH_3$ complex has relatively high energy. Table 4 shows that the reaction $(NH_3)AgCN \rightarrow 1/2\ [Ag(NH_3)_2]^+ + 1/2\ [Ag(CN)_2]^-$ is exothermic with a significant energy difference, i.e., 32.1 (B3LYP) and 32.8 (BP86) kcal/mol. Thus, the newly prepared Ag(CN)$NH_3$ complex does not display absolute thermodynamical stability. This may explain why Ag(CN)$NH_3$ was prepared many years after $[Ag(NH_3)_2]^+$ and $[Ag(CN)_2]^-$ became common chemical reagents.

It is also known that the one-dimensional linear, polymeric chain structures are common for AgCN and the other group 11 metal cyanides.[36,37] The geometry of the chain structure −Ag−CN−Ag−CN− was determined by the neutron diffraction experiment in 2002.[38] Indeed, our theoretical results have confirmed that the linear AgCN−AgCN dimer has a decomposition energy (to two AgCN's) of ∼37 kcal/mol (Table 4). Similarly, our theoretical results show that the chain structure $(NH_3)AgCN{-}AgCN$ has a lower energy than separated $(NH_3)AgCN$ and AgCN by about 42 kcal/mol (Table 4), and thus the monomer Ag(CN)$NH_3$ is only a metastable structure.

**NC−Ag−$NF_3$.** When the $NH_3$ ligand in Ag(CN)$NH_3$ is replaced by the more electronegative $NF_3$ ligand, the linear $C_{3v}$ Ag(CN)$NF_3$ complex is also predicted as a genuine minimum (Figure 1 and Table 5). However, there are some important differences between Ag(CN)$NH_3$ and Ag(CN)$NF_3$. To begin, the $NF_3$ fragment in Ag(CN)$NF_3$ is geometrically somewhat different from the isolated $NF_3$. The N−F distances increase by 0.014 Å, and the F−N−F bond angles increase by more than 1° (Tables 2 and 5). The Ag−$NF_3$ bond distance (2.294 Å) in Ag(CN)$NF_3$ is longer than the Ag−$NH_3$ bond distance (2.186 Å) in Ag(CN)$NH_3$ (Table 1) by more than 0.1 Å, suggesting a weaker interaction between

***Table 3.*** B3LYP Natural Atomic Charges (Q) and Wiberg Bond Indices (WBI) of Selected Bonds of Ag(CN)$XY_3$ (X = N and P, Y = H and F)

| | $Q_{Ag}$ | $Q_C$ | $Q_N$ | $Q_X$ | $Q_Y$ | $Q_{AgCN}$ | $Q_{XY3}$ | $WBI_{Ag\text{-}XY3}$ | $WBI_{Ag\text{-}CN}$ |
|---|---|---|---|---|---|---|---|---|---|
| Ag(CN)$NH_3$ | 0.54 | −0.16 | −0.50 | −1.12 | 0.41 | −0.12 | 0.12 | 0.20 | 0.58 |
| Ag(CN)$NF_3$ | 0.57 | −0.17 | −0.47 | 0.59 | −0.17 | −0.06 | 0.06 | 0.16 | 0.60 |
| Ag(CN)$PH_3$ | 0.50 | −0.18 | −0.49 | 0.08 | 0.03 | −0.17 | 0.17 | 0.36 | 0.53 |
| Ag(CN)$PF_3$ | 0.45 | −0.20 | −0.46 | 1.85 | −0.55 | −0.21 | 0.21 | 0.45 | 0.52 |
| Ag(CN) | 0.67 | −0.20 | −0.47 | | | | | | 0.60 |

***Table 5.*** Geometrical Parameters for M(CN)$NF_3$ (M = Cu, Ag, Au)[a]

| species | | B3LYP Cu | B3LYP Ag | B3LYP Au | BP86 Cu | BP86 Ag | BP86 Au |
|---|---|---|---|---|---|---|---|
| M(CN)($NF_3$) | M−C | 1.842 | 2.021 | 1.953 | 1.832 | 1.998 | 1.944 |
| | C−N | 1.173 | 1.172 | 1.171 | 1.186 | 1.185 | 1.184 |
| | M−N | 1.978 | 2.294 | 2.171 | 1.907 | 2.210 | 2.112 |
| | N−F | 1.376 | 1.372 | 1.372 | 1.407 | 1.397 | 1.400 |
| | ∠F−N−F | 103.0 | 103.2 | 103.2 | 102.4 | 102.9 | 102.7 |

[a] Bond distances in Å, bond angles in degrees.

***Table 6.*** Dissociation Energies (in kcal/mol) for M(CN)$XY_3$ and M(Br)$XY_3$ (M = Cu, Ag, and Au; X = N and P; Y = H and F)

| | B3LYP Cu | B3LYP Ag | B3LYP Au | BP86 Cu | BP86 Ag | BP86 Au |
|---|---|---|---|---|---|---|
| M(CN)($NH_3$) → M(CN) + $NH_3$ | 41.7 | 33.4 | 43.4 | 43.2 | 35.3 | 44.5 |
| M(CN)($NF_3$) → M(CN) + $NF_3$ | 15.5 | 8.6 | 15.1 | 18.4 | 10.0 | 16.8 |
| M(CN)($PH_3$) → M(CN) + $PH_3$ | 33.5 | 29.0 | 41.4 | 36.1 | 31.7 | 43.2 |
| M(CN)($PF_3$) → M(CN) + $PF_3$ | 24.0 | 18.2 | 31.6 | 27.9 | 21.9 | 34.6 |

the $NF_3$ fragment and the AgCN fragment. On the basis of the NBO analysis, the Wiberg Ag−N bond index (0.16) for Ag(CN)$NF_3$ is somewhat smaller than that (0.20) for Ag(CN)$NH_3$. More importantly, the natural charge for the N atom in $NF_3$ ($Q_X$) is no longer negative, and the charges for the AgCN and $NF_3$ fragments are only −0.06 and +0.06, respectively, which are only half of those for Ag(CN)$NH_3$ (Table 3), suggesting that the ionic attraction between the two parts is much weaker. Accordingly, the dissociation energy for Ag(CN)$NF_3$ to lose $NF_3$ is only 8.6 (B3LYP) or 10.0 (BP86) kcal/mol (Table 6), compared with the substantial value (∼34 kcal/mol) for Ag(CN)$NH_3$ to lose $NH_3$. This small dissociation energy may make it difficult to synthesize Ag(CN)$NF_3$ in a manner analogous to that used for Ag(CN)-$NH_3$.

**NC−M−$NH_3$ and NC−M−$NF_3$ (M = Cu, Au).** Although CuCN·$NH_3$ has long been a known compound, the crystal structure for the CuCN·$NH_3$ complex was reported to have a copper coordination number of four,[39] which is quite different from the presently considered linear ($NH_3$)AgCN complex. For gold, no similar complexes have been reported. In 2007, Mishra used the *ab initio* methods to predict the linear structures for $XAuY^-$ (X, Y = Cl, Br, and I).[40]

In the present study, like Ag(CN)$NH_3$, the linear asymmetric Cu(CN)$NH_3$ and Au(CN)$NH_3$ structures (with $C_{3v}$ symmetry, Figure 1) are predicted to be genuine minima, and their geometry parameters are reported in Table 1. The $NH_3$ fragment in Cu(CN)$NH_3$ or in Au(CN)$NH_3$ has almost the same geometry as that in Ag(CN)$NH_3$, with the change less than 0.001 Å for the N−H distance and 0.3° for the H−N−H angle (B3LYP method). This suggests that the interactions between the $NH_3$ ligand and CuCN and AuCN are also mainly the classical electrostatic attraction.

Energetically, the Cu(CN)$NH_3$ and Au(CN)$NH_3$ complexes have dissociation energies for the reaction M(CN)($NH_3$) → M(CN) + $NH_3$ (M = Cu and Au) that are even larger than that predicted for Ag(CN)$NH_3$. The dissociation energy for Cu(CN)$NH_3$ is 41.7 (B3LYP) or 43.2 (BP86) kcal/mol, and that for Au(CN)$NH_3$ is 43.4 (B3LYP) or 44.5 (BP86) kcal/mol (Table 6). These values are also comparable with the DFT dissociation energy predicted by Frenking et al. for the related metal−carbene bond of XM−L (X = F−I; M = Cu, Ag, Au; L = imidazol-2-ylidene).[41] The energy decomposition analysis (EDA) performed by Frenking et al. shows that the metal−carbene bonds are mainly held together by classical electrostatic attraction (>65%), consistent with our above-noted descriptions of the M−$NH_3$ bonds in the NC−M−$NH_3$ (M = Cu, Ag, Au) complexes. Since the copper and gold analogues of the simple linear asymmetric complexes of Ag(CN)$NH_3$ have properties similar to those of Ag(CN)$NH_3$, these analogues are likely to be synthesizable.

The $NF_3$ analogous complexes Cu(CN)$NF_3$ and Au(CN)-$NF_3$ have also been studied here, and the results are listed in Table 5. The linear structures are predicted to be genuine minima on their potential hypersurfaces. Like the situation in Cu(CN)$NH_3$ and Au(CN)$NH_3$, however, the analogous complexes Cu(CN)$NF_3$ and Au(CN)$NF_3$ suggest classical electrostatic attractions between $NF_3$ and MCN. Table 6 compares the dissociation energies for M(CN)$NF_3$ (M = Cu, Ag, Au) with those predicted for M(CN)$NH_3$. The dissociation energies of Cu(CN)$NF_3$ and Au(CN)$NF_3$ for losing $NF_3$ are about 15 kcal/mol (B3LYP, Table 6), substantially smaller compared with the corresponding M(CN)$NH_3$ complexes (>40 kcal/mol, Table 6), but not as small as the dissociation energy (∼9 kcal/mol, Table 6) for Ag(CN)$NF_3$ to lose $NF_3$.

***Table 7.*** Geometrical Parameters for M(CN)$PH_3$ and M(CN)$PH_3$ (M = Cu, Ag, Au)[a]

| species | | B3LYP Cu | B3LYP Ag | B3LYP Au | BP86 Cu | BP86 Ag | BP86 Au |
|---|---|---|---|---|---|---|---|
| M(CN)($PH_3$) | M−C | 1.867 | 2.046 | 2.000 | 1.852 | 2.026 | 1.987 |
| | C−N | 1.174 | 1.173 | 1.172 | 1.187 | 1.186 | 1.185 |
| | M−P | 2.226 | 2.409 | 2.333 | 2.197 | 2.371 | 2.315 |
| | P−H | 1.410 | 1.409 | 1.408 | 1.423 | 1.421 | 1.421 |
| | ∠H−P−H | 99.1 | 99.0 | 100.0 | 98.7 | 98.5 | 99.5 |
| M(CN)($PF_3$) | M−C | 1.864 | 2.040 | 1.998 | 1.852 | 2.022 | 1.987 |
| | C−N | 1.173 | 1.173 | 1.171 | 1.186 | 1.185 | 1.184 |
| | M−P | 2.186 | 2.382 | 2.282 | 2.157 | 2.339 | 2.266 |
| | P−F | 1.567 | 1.568 | 1.564 | 1.584 | 1.585 | 1.580 |
| | ∠F−P−F | 100.0 | 100.0 | 100.2 | 100.0 | 100.1 | 100.2 |

[a] Bond distances in Å, bond angles in degrees.

**M(CN)$PH_3$ and M(CN)$PF_3$ (M = Cu, Ag, Au).** We extended our studies to the phosphorus analogues (i.e., to replace $NH_3$ and $NF_3$ with $PH_3$ and $PF_3$). The two DFT methods predict the similar linear $C_{3v}$ minima for Ag(CN)$PH_3$ and Ag(CN)$PF_3$, and their structures are reported in Table 7. The geometry parameters for the $PH_3$ and $PF_3$ ligands in Ag(CN)$PH_3$ and Ag(CN)$PF_3$ (Table 7) are slightly different from the isolated $PH_3$ and $PF_3$ species (Table 2). The dissociation energy for Ag(CN)$PH_3$ is substantial, i.e., 29.0 (B3LYP) or 31.7 (BP86) kcal/mol (Table 6), which is comparable to that for Ag(CN)$NH_3$. The dissociation energy for Ag(CN)$PF_3$ is smaller, i.e., 18.2 (B3LYP) or 21.9 (BP86) kcal/mol, but not as small as that for Ag(CN)$NF_3$ (9 kcal/mol, Table 6). Nevertheless, these complexes are favorable with respect to the loss of the $PH_3$ and $PF_3$ ligands. The Ag−P distances in Ag(CN)$PH_3$ are slightly (∼0.03 Å) longer than that in Ag(CN)$PF_3$, consistent with the corresponding Wiberg bond indices (Table 3).

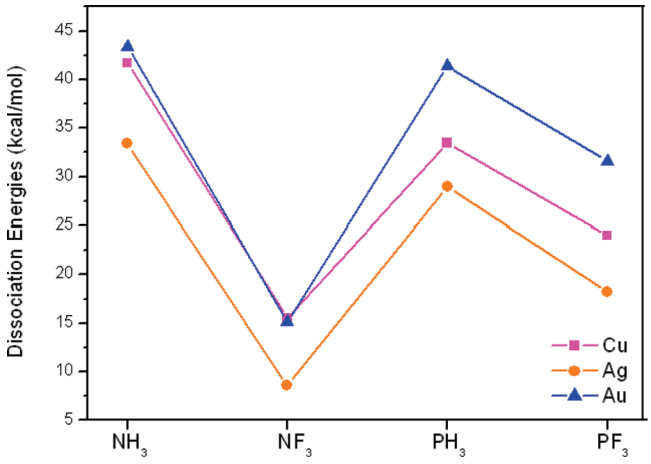



***Figure 3.*** The dissociation energies of M(CN)($XY_3$) (M = Cu, Ag, Au; X = N, P; Y = H, F).

It may be seen from Table 3 that the positive charges on Ag in Ag(CN)$PY_3$ are slightly less than those for Ag(CN)-$NY_3$, and the Wiberg bond indices of Ag−P in Ag(CN)$PY_3$ are larger than those of Ag−N in Ag(CN)$NY_3$. For Ag(C-N)$PF_3$, the WBI(Ag−P) of 0.45 is the largest among the four Ag(CN)$XY_3$ (X = N or P; Y = H or F) species, and the P atom has the largest positive charge of +1.85 (Table 3). These results show that the bonds between the Ag and P atoms in Ag(CN)$PY_3$ are less ionic but more covalent than the bonds between the Ag and N atoms in Ag(CN)$NY_3$.

The Cu and Au analogues M(CN)$PH_3$ and M(CN)$PF_3$ (M = Cu, Au) also have the linear asymmetric structure, and their geometries are reported in Table 7. The $PH_3$ and $PF_3$ ligands in the Cu and Au analogues have almost the same geometries as those in Ag(CN)$PH_3$ or Ag(CN)$PF_3$. The dissociation energies for losing the $PH_3$ and $PF_3$ ligands are reported in Table 6, and these values are comparable with (or even larger than) those of Ag(CN)$PH_3$ or Ag(CN)$PF_3$. Therefore, all of these complexes are energetically viable.

Figure 3 compares the B3LYP bond dissociation energies (BDE) for the 12 M(CN)$XY_3$ species. The complexes containing the $NH_3$ ligand, i.e., M(CN)$NH_3$ for all three metals, have the largest M−N bond dissociation energies (>33 kcal/mol), while the $NF_3$-containing complexes have the smallest BDEs (as small as 8.6 kcal/mol for Ag(CN)$NF_3$). This is consistent with the NBO analysis mentioned above. The ionic interaction between M and N is believed to play an important role in stabilizing these complexes. The dissociation energies for M(CN)$PH_3$ are larger than those for M(CN)$PF_3$, but the difference is not as large as those between the M(CN)$NH_3$ and M(CN)$NF_3$ species. Thus, the BDE values for the complexes containing the different ligands are in the order $NH_3 > PH_3 > PF_3 > NF_3$. Comparing complexes among the three different coinage metals, the theoretical BDE values show the order Au > Cu > Ag. This is in agreement with the general trend for the bond dissociation energies of the first, second, and third transition metal rows.[42] On the basis of energy considerations, the linear asymmetric complexes for Cu and Au should be even more favorable species than their Ag analogues.

**Ag(Br)$NH_3$ and Analogues.** In Chippindale et al.'s 2008 paper,[13] the other critical new structure synthesized was the

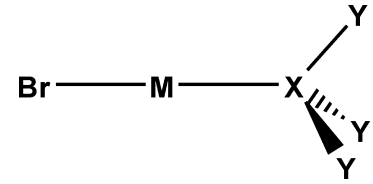



***Figure 4.*** Qualitative structures of BrM($XY_3$) (M = Cu, Ag, Au; X = N, P; Y = H, F).

analogous asymmetric complex of Ag(I), namely, Ag-(Br)$NH_3$. The molecular units in the crystal slightly distort from linearity with an experimental Br−Ag−N bond angle of 165.01(12)°. However, in the present research, the theoretically optimized geometry for the isolated Ag(Br)$NH_3$ remains a linear structure with $C_{3v}$ symmetry (Figure 4 and Table 8). Chippindale et al. have suggested that the slight distortion from the linear geometry for the crystal Ag(Br)$NH_3$ is caused by a weak intermolecular interaction, i.e., between a silver atom and a bromine atom in a neighboring molecule (with 2.971 Å being the intermolecular Ag···Br distance).[13] Our predicted Ag−N distance (2.204 Å by B3LYP or 2.171 Å by BP86) is close the crystal Ag−N distance 2.192 Å. Our predicted Ag−Br distance (2.435 Å by B3LYP and 2.413 Å by BP86) is somewhat shorter than the crystal Ag−Br distance (2.536 Å),[13] and the Ag−Br distance (2.49 Å) in the $[AgBr_2]^-$ anion.[43] The dissociation energy for Ag(Br)$NH_3$ to lose $NH_3$ is predicted to be 29.6 (B3LYP) or 32.0 (BP86) kcal/mol (Table 10), which is ∼2 kcal/mol smaller than that for Ag(CN)$NH_3$. This may explain why the Ag(Br)$NH_3$ complex, while also successfully prepared experimentally, decomposes at lower temperature than Ag(CN)$NH_3$.

Like for the M(CN)$XY_3$ analogues, we have investigated a total of 12 M(Br)$XY_3$ analogues (M = Cu, Ag, Au; $XY_3$ = $NH_3$, $NF_3$, $PH_3$, $PF_3$) with the same methods. All of these complexes are predicted to be linear in the heavy atoms, with $C_{3v}$ symmetry, and their geometry parameters are reported in Table 8. The M−X distances and the geometry parameters for $XY_3$ in the 12 M(Br)$XY_3$ complexes are very close to the corresponding geometry parameters in M(CN)$XY_3$ (Tables 1, 5, and 7). Comparing the geometries of the M(Br)$XY_3$ complexes with the free MBr and $XY_3$, the trend of the changes of the corresponding geometrical parameters is similar to that for the M(CN)$XY_3$ species.

The natural bond orbital (NBO) analyses were also carried out for the Ag(Br)$XY_3$ complexes (Table 9). It is interesting to compare the natural atomic charges and Wiberg bond indices (WBI) for the Ag(Br)$XY_3$ complexes with those for their CN analogues (Table 3). These two tables are remarkably similar with a deviation of less than 0.04 for natural atomic charges and 0.07 for WBIs, if CN is considered the analogue of Br. Similar to the above discussion for the Ag(CN)$XY_3$ complexes (Table 3), there is primarily ionic bonding between AgBr and $XY_3$. The Ag−$NH_3$ interaction is stronger than that for Ag−$NF_3$, since the latter has significantly smaller atomic charges for $Q_X$, $Q_Y$, $Q_{AgBr}$, and $Q_{XY3}$ (Table 9). For the complexes with $PH_3$ and $PF_3$ ligands, the absolute values of the fragment natural charges $Q_{AgBr}$ and $Q_{XY3}$ are substantial (>0.14), suggesting strong bonding. However, the larger

***Table 8.*** Geometrical Parameters for M(Br)$XY_3$ (M = Cu, Ag, Au; X = N, P; Y = H, F)[a]

| species | | B3LYP Cu | B3LYP Ag | B3LYP Au | BP86 Cu | BP86 Ag | BP86 Au |
|---|---|---|---|---|---|---|---|
| M(Br)($NH_3$) | M−Br | 2.229 | 2.435 | 2.411 | 2.216 | 2.413 | 2.398 |
| | M−N | 1.951 | 2.204 | 2.136 | 1.932 | 2.171 | 2.118 |
| | N−H | 1.023 | 1.021 | 1.022 | 1.031 | 1.029 | 1.030 |
| | ∠H−N−H | 107.2 | 107.4 | 107.8 | 107.0 | 107.2 | 107.8 |
| M(Br)($NF_3$) | M−Br | 2.210 | 2.424 | 2.390 | 2.199 | 2.399 | 2.376 |
| | M−N | 1.956 | 2.326 | 2.150 | 1.872 | 2.201 | 2.063 |
| | N−F | 1.379 | 1.374 | 1.377 | 1.417 | 1.402 | 1.410 |
| | ∠F−N−F | 102.7 | 103.0 | 102.7 | 101.9 | 102.6 | 102.0 |
| M(Br)($PH_3$) | M−Br | 2.237 | 2.441 | 2.431 | 2.224 | 2.422 | 2.419 |
| | M−P | 2.197 | 2.397 | 2.292 | 2.161 | 2.350 | 2.268 |
| | P−H | 1.411 | 1.410 | 1.409 | 1.424 | 1.422 | 1.422 |
| | ∠H−P−H | 98.8 | 98.7 | 99.8 | 98.3 | 98.2 | 99.3 |
| M(Br)($PF_3$) | M−Br | 2.221 | 2.421 | 2.411 | 2.211 | 2.402 | 2.399 |
| | M−P | 2.156 | 2.364 | 2.242 | 2.122 | 2.312 | 2.222 |
| | P−F | 1.570 | 1.570 | 1.567 | 1.588 | 1.588 | 1.584 |
| | ∠F−P−F | 99.6 | 99.7 | 99.8 | 99.5 | 99.7 | 99.8 |
| MBr | M−Br | 2.216 | 2.449 | 2.412 | 2.198 | 2.428 | 2.394 |

[a] Bond distances in Å, bond angles in degrees.

***Table 9.*** B3LYP Natural Atomic Charges ($Q$) and Wiberg Bond Indices (WBI) of Selected Bonds of Ag(Br)$XY_3$ (X = N and P, Y = H and F)

| | $Q_{Ag}$ | $Q_{Br}$ | $Q_X$ | $Q_Y$ | $Q_{AgBr}$ | $Q_{XY3}$ | $WBI_{Ag\text{-}XY3}$ | $WBI_{Ag-Br}$ |
|---|---|---|---|---|---|---|---|---|
| Ag(Br)$NH_3$ | 0.58 | −0.68 | −1.13 | 0.41 | −0.10 | 0.10 | 0.20 | 0.51 |
| Ag(Br)$NF_3$ | 0.60 | −0.65 | 0.59 | −0.18 | −0.04 | 0.04 | 0.15 | 0.57 |
| Ag(Br)$PH_3$ | 0.53 | −0.68 | 0.06 | 0.03 | −0.14 | 0.14 | 0.37 | 0.49 |
| Ag(Br)$PF_3$ | 0.48 | −0.65 | 1.83 | −0.55 | −0.17 | 0.17 | 0.46 | 0.50 |
| AgBr | 0.66 | | | | | | | 0.60 |

$WBI_{Ag\text{-}XY3}$ values (>0.37) suggest that the bonding between AgBr and $PY_3$ possesses more covalent character.

The bond dissociation energies (BDE) for all 12 M(Br)$XY_3$ complexes are reported in Table 10 and displayed in Figure 5. Compared with those in Table 6 and Figure 3 for M(CN)$XY_3$, the BDEs of M(Br)$XY_3$ (Figure 5) have a very similar pattern. As for the three different coinage metals, Figure 5 shows that the theoretical BDE values for M(Br)$XY_3$ follow the irregular order Au > Cu > Ag, as was the case for the M(CN)$XY_3$ complexes. On the basis of energy considerations, the linear asymmetric M(Br)$XY_3$ complexes for Cu and Au have good prospects for preparation.

## Summary

The present research shows that Ag(CN)$NH_3$ has a $C_{3v}$ linear asymmetric structure, with theoretical bond distances in good agreement with the experimental crystal structure. Our structure for Ag(Br)$NH_3$ is linear in the three heavy atoms, while the experimental structure is slightly bent (165.0°). The difference between the theory and experiment has been rationalized in terms of the intermolecular interactions.[13]

The bond dissociation energies (BDE) are shown in Tables 6 and 10, as well as in Figures 3 and 5. For Cu and Ag, the dissociation energies fall in the order $NH_3 > PH_3 > PF_3 > NF_3$. However, for Au, the dissociation energies for Au(CN)$XY_3$ are in the order $NH_3 \sim PH_3 > PF_3 > NF_3$, and those for Au(Br)$XY_3$ become different, i.e., $PH_3 > NH_3 \sim PF_3 > NF_3$ (Tables 6 and 10). Thus, it is understandable that the dissociation energies were reported in the order $PH_3 > PF_3 > NH_3 > NF_3$ for ClAu−L reported by Pyykkö et al. at the BP86/cc-pVDZ level of theory.[44]

***Table 10.*** Dissociation Energies (in kcal/mol) for M(Br)L (M = Cu, Ag, and Au; L = $NH_3$, $NF_3$, $PH_3$, $PF_3$)

| | B3LYP Cu | B3LYP Ag | B3LYP Au | BP86 Cu | BP86 Ag | BP86 Au |
|---|---|---|---|---|---|---|
| M(Br)($NH_3$) → M(Br) + $NH_3$ | 39.1 | 29.6 | 39.3 | 41.7 | 32.0 | 42.2 |
| M(Br)($NF_3$) → M(Br) + $NF_3$ | 15.0 | 6.9 | 13.3 | 20.3 | 9.1 | 17.7 |
| M(Br)($PH_3$) → M(Br) + $PH_3$ | 33.6 | 27.2 | 42.3 | 38.0 | 31.2 | 46.8 |
| M(Br)($PF_3$) → M(Br) + $PF_3$ | 25.7 | 17.5 | 34.4 | 31.8 | 22.9 | 40.3 |

The NBO analysis shows that the interaction between AgCN and $NH_3$ (or AgBr and $NH_3$) is essentially ionic, and this is in agreement with the EDA analysis for related carbenes by Frenking et al.[41] The dissociation energy of Ag(CN)$NH_3$ for losing $NH_3$ is predicted to be substantial (>30 kcal/mol), suggesting an energetically viable species.

We have also studied a series of analogues of Ag(CN)$NH_3$ and Ag(Br)$NH_3$, i.e., XML (X = CN, Br; M = Cu, Ag, Au; L = $NH_3$, $NF_3$, $PH_3$, $PF_3$), most of which are experimentally unknown. All of these analogous compounds are predicted to have similar $C_{3v}$ linear asymmetric minima. Except for the $NF_3$-containing complexes, the dissociation energies for these analogous complexes are substantial, and they should have synthetic potential, as recently discovered experimentally for Ag(CN)$NH_3$ and Ag(Br)$NH_3$.

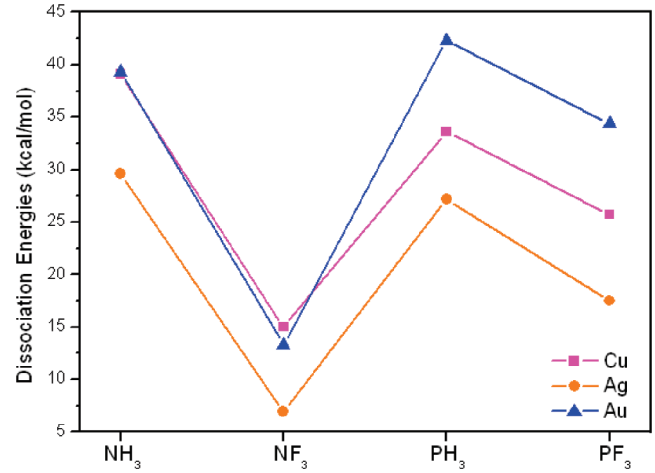



***Figure 5.*** Dissociation energies of M(Br)($XY_3$) (M = Cu, Ag, Au; X = N, P; Y = H, F).

**Acknowledgment.** This research was supported in the USA by the National Science Foundation, Grants CHE-0749868 and CHE-0716718; and in China the National Natural Science Foundation (20802093), Excellent talents training Fund of Beijing (2010D009011000003), the research fund for the doctoral program of higher education (20060007030 and 20070533142), the scientific research fund of state key laboratory of explosion science and technology, and Excellent Young Scholars Research Fund of Beijing Institute of Technology (2008Y0206).

## References

(1) Earnshaw, A., Greenwood, N. N. In *Chemistry of the Elements*; Elsevier: Oxford, U. K., 1997; pp 1194−1197.

(2) Pyykkö, P. *Chem. Soc. Revs.* **2008**, *37*, 1967.

(3) Khairallah, G. N.; O'Hair, R. A. J.; Bruce, M. I. *J. Chem. Soc., Dalton. Trans.* **1998**, 3935.

(4) Larsen, A. O.; Leu, W.; Oberhuber, C. N.; Campbell, J. E.; Hoveyda, A. H. *J. Am. Chem. Soc.* **2004**, *126*, 11130.

(5) Van Veldhuizen, J. J.; Campbell, J. E.; Giudici, R. E.; Hoveyda, A. H. *J. Am. Chem. Soc.* **2005**, *127*, 6877.

(6) Tonner, R.; Heydenrych, G.; Frenking, G. *Chem. Asian J.* **2007**, *2*, 1555.

(7) Chung, M. C. *Bull. Korean Chem. Soc.* **2002**, *23*, 921.

(8) Tulloch, A. A. D.; Danopoulos, A. A.; Winston, S.; Kleinhenz, S.; Eastham, G. *J. Chem. Soc., Dalton Trans.* **2000**, 4499.

(9) Simons, R. S.; Custer, P.; Tessier, C. A.; Youngs, W. J. *Organometallics* **2003**, *22*, 1979.

(10) Chen, W.; Wu, B.; Matsumoto, K. *J. Organomet. Chem.* **2002**, *654*, 233.

(11) Pytkowicz, J.; Roland, S.; Mangeney, P. *J. Organomet. Chem.* **2001**, *631*, 157.

(12) César, V.; Bellemin-Laponnaz, S.; Gade, L. H. *Organometallics* **2002**, *21*, 5204.

(13) Chippindale, A. M.; Head, L. E.; Hibble, S. J. *Chem. Commun.* **2008**, 3010.

(14) Garrison, J. C.; Youngs, W. J. *Chem. Rev.* **2005**, *105*, 3978.

(15) Hartshorn, C. M.; Steel, P. J. *J. Chem. Soc. Dalton. Trans.* **1998**, 3935.

(16) Bowyer, P. K.; Porter, K. A.; David, R. A.; Willis, A. C.; Wild, S. B. *J. Chem. Soc. Chem. Commun.* **1998**, 1153.

(17) Sailaja, S.; Rajasekharan, M. V. *Inorg. Chem.* **2000**, *39*, 4586.

(18) Liao, S.; Su, C. Y.; Yeung, C. H.; Xu, A. W.; Zhang, H. X.; Liu, H. Q. *Inorg. Chem. Commun.* **2000**, *3*, 405.

(19) Zhao, Y. J.; Hong, M. C.; Liang, Y. C.; Su, W. P.; Cao, R.; Zhou, Z. Y.; Chan, A. S. C. *Polyhedron* **2001**, *20*, 2619.

(20) Mikhailov, V. A.; Tew, D. P.; Walker, N. R.; Legon, A. C. *Chem. Phys. Lett.* **2010**, *499*, 16.

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

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

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

(24) Perdew, P. J. *Phys. Rev. B* **1986**, *33*, 8822.

(25) Andrae, D.; Haussermann, U.; Dolg, M.; Stoll, H.; Press, H. *Theor. Chim. Acta* **1990**, *77*, 123.

(26) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(27) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.

(28) Dunning, T. H.; Hay, P. J. In *Methods of Electronic Structure Theory*; Schaefer, H. F., Ed.; Plenum: New York, 1977; pp 1−27.

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

(30) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.

(32) Grotjahn, D. B.; Brewster, M. A.; Ziurys, L. M. *J. Am. Chem. Soc.* **2002**, *124*, 5895.

(33) Okabayashi, T.; Okabayashi, E. Y.; Koto, F.; Ishida, T.; Tanimoto, M. *J. Am. Chem. Soc.* **2009**, *131*, 11712.

(34) Zaleski-Ejgierd, P.; Patzschke, M.; Pyykkö, P. *J. Chem. Phys.* **2008**, *128*, 224303.

(35) Paul, A.; Yamaguchi, Y.; Schaefer, H. F. *J. Chem. Phys.* **2007**, *127*, 154324.

(36) Kroeker, S.; Wasylishen, R. E.; Hanna, J. V. *J. Am. Chem. Soc.* **1999**, *121*, 1582.

(37) Bowmaker, G. A.; Kennedy, B. J.; Reid, J. C. *Inorg. Chem.* **1998**, *37*, 3968.

(38) Hibble, S. J.; Cheyne, S. M.; Hannon, A. C.; Eversfield, S. G. *Inorg. Chem.* **2002**, *41*, 1042.

(39) Cromer, D. T.; Larson, A. C.; Roof, R. B. *Acta Crystallogr.* **1965**, *19*, 192.

(40) Mishra, S. *J. Phys. Chem. A* **2007**, *111*, 9164.

(41) Nemcsok, D.; Wichmann, K.; Frenking, G. *Organometallics* **2004**, *23*, 3640.

(42) Frenking, G.; Antes, I.; Boehme, M.; Dapprich, S.; Ehlers, A.; Jonas, V.; Neuhaus, A.; Otto, M.; Stegmann, R.; Veldkamp, A.; Vyboyshchikov, S. F., In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1996; Vol. 8, pp 81−83.

(43) Rabilloud, F.; Spiegelmann, F.; Heully, J. L. *J. Chem. Phys.* **1999**, *111*, 8925.

(44) Pyykkö, P.; Runeberg, N. *Chem. Asian J.* **2006**, *1*, 623.

CT100664Q

JCTC Journal of Chemical Theory and Computation

# Role of Cation Polarization in *holo*- and *hemi*-Directed $[Pb(H_2O)_n]^{2+}$ Complexes and Development of a $Pb^{2+}$ Polarizable Force Field

Mike Devereux,*,† Marie-Céline van Severen,‡ Olivier Parisel,‡ Jean-Philip Piquemal,*,‡,§ and Nohad Gresh*,†

*Université Paris Descartes, Laboratoire de Chimie et Biochimie Pharmacologiques et Toxicologiques, UMR 8601 CNRS, UFR Biomédicale, 45 rue des Saints-Pères, 75270 Paris Cedex06, France; UPMC, Université Paris 06, UMR 7616, Laboratoire de Chimie Théorique, Case Courrier 137, 4 Place Jussieu, F-75005 Paris, France; and CNRS, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005 Paris, France*

Received July 19, 2010

**Abstract:** Reduced Variational Space (RVS) calculations are reported that afford insight into the energetic origins of the *hemi*- and *holo*-directing behavior of $[Pb(H_2O)_n]^{2+}$ complexes. It is shown that the distribution of ligands around the $Pb^{2+}$ center arises from a delicate balance between the first-order Coulomb plus exchange-repulsion energy that favors *holo*-directionality, and the second-order charge transfer plus polarization term that favors *hemi*-directionality. It is additionally demonstrated that the pseudopotential/basis set combination used to study such complexes should be carefully selected, as artifacts can arise when using large-core pseudopotentials. Finally, based on these findings, we introduce a new SIBFA force field parametrization for $Pb^{2+}$. Results yield close agreement with ab initio complexation energies in a series of $[Pb(H_2O)_n]^{2+}$ complexes and successfully encapsulate the *hemi*- and *holo*-directing properties. SIBFA thus appears to be the first classical force field to be able to model the *holo*-/*hemi*-directed transition within Pb complexes, avoiding the need for explicit wave function treatment and consequently providing the opportunity to deal with large leaded systems of biological interest.

## Introduction

As lead is an abundant metal that is easily extracted and has a low melting point and high malleability, it has been widely used for everything from making cooking utensils, paints, or water pipes to electrochemical cells, and as an additive to gasoline in internal combustion engines. As a result, it has become widely dispersed in the environment,[1] and due to the high toxicity associated with $Pb^{2+}$, poses a significant threat to human health.[2] Pb accumulated in the body causes lead poisoning, or saturnism, an intoxication that is especially severe among children[3] and threatens many people in developing nations where contact with lead-contaminated soils and drinking water is most common.[4]

Lead is easily oxidized to form $Pb^{2+}$ aquacations. $Pb^{2+}$ then competes with and displaces native cations such as $Zn^{2+}$ in important proteins such as Aminolevulinic Acid Dehydratase (ALAD),[5−9] inhibiting normal protein function. Explaining the affinity and structural changes associated with binding of $Pb^{2+}$ in complex biological environments and designing chelating agents that are able to selectively extract $Pb^{2+}$ in preference to other metal cations is a particularly challenging task.[10] The high atomic number of $Pb^{2+}$ ($Z =$

* To whom correspondence should be addressed. E-mail: mike.devereux@parisdescartes.fr; nohad.gresh@parisdescartes.fr; jpp@lct.jussieu.fr.

† Université Paris Descartes, U648 INSERM, UFR Biomédicale.

‡ Université Pierre et Marie Curie—Paris 06, UMR 7616, Laboratoire de Chimie Théorique.

§ CNRS, UMR 7616, Laboratoire de Chimie Théorique.

10.1021/ct1004005 
Published on Web 11/30/2010

82) makes full-electron, relativistic ab initio calculations prohibitive for all but the smallest complexes. There is also an unusual tendency of $Pb^{2+}$ complexes to switch between even spacing of ligands around the metal center in a so-called *holo*-directed conformation, and a second class of structures where ligands are directed to one side of the metal in a sterically crowded *hemi*-directed conformation. The preference for *holo*- or *hemi*-directionality depends on many parameters, such as the number of coordinating ligands and resulting steric crowding, the ligand flexibility, and additional repulsion between ligands that carry a formal charge.[11] The reason $Pb^{2+}$ exhibits this unusual behavior is often attributed to a sterically active lone-pair,[12,13] making it difficult to model using traditional classical force fields. The availability of a reliable molecular mechanics approach would allow studies of $Pb^{2+}$ binding with greater conformational sampling and in more complex environments than those currently accessible by means of electronic structure methods.

As force fields typically combine separate energetic terms to estimate the total interaction energy between moieties, a deeper understanding of the energetic contributions favoring *holo* and *hemi* orientation in different complexes is an important first step toward parametrizing a reliable force field. One methodology that can be applied in this context is the Reduced Variational Space (RVS) scheme of Stevens and Fink.[14] RVS decomposes the total ab initio interaction energy between fragments into first-order Coulomb and Pauli-repulsion terms, and second-order polarization and charge-transfer components. All of these contributions are accounted for using approximate expressions in the SIBFA (Sum of Interactions Between Fragments Ab initio computed) polarizable force field.[15−19] As a result, RVS has been successfully used for the energetic decomposition of various systems including $Zn^{2+}$-complexes,[20] water clusters[21] and hard and soft metal cations,[22] to allow parametrization of SIBFA and for subsequent evaluation of force field accuracy.

SIBFA is a detailed, fragment-based force field that has been widely applied to proteins and organometallic systems, from small complexes to proteins/metalloproteins[23−26] (see ref 19 for a detailed review article). The force field energetic terms explicitly account for anisotropic fragment polarization, repulsion, and charge transfer, as well as for distributed multipole moments centered at both nuclear positions and bond barycenters. While computationally more expensive than many popular classical force fields such as CHARMM[27] and AMBER,[28] SIBFA yields molecular structures and energies that are generally in close agreement with ab initio data while maintaining a much lower computational overhead than would be required for a full ab initio computation. Although the interest in polarizable, anisotropic force fields has recently grown,[29−31] SIBFA remains one of the most developed and widely applied. Parameters exist for a wide range of organic compounds and closed-shell organometallic systems that contain metals such as Mg(II), Ca(II), Zn(II), and Cd(II). A recent extension, "SIBFA-LF", includes ligand field effects by means of the angular overlap model[32,33] (AOM) and has allowed extension of the range of applications to transition metal cations with partially filled *d*-shells, such as Cu(II).[15] In addition, short-range energetic corrections allow SIBFA to describe the formation of ligand−metal complexes without resorting to separate bonded and non-bonded parameters. A single, consistent parameter set can thus be used to explore for example binding, dissociation and relative energies of different tautomers and conformers. The goal of the current work is to use information gained about the formation of $Pb^{2+}$-complexes from RVS calculations to extend the range of application of SIBFA to $Pb^{2+}$ compounds, offering a capability for molecular mechanics to model the *holo*- to *hemi*-directing transition.

## Computational Procedures

**Electronic Structure Calculations.** To explore the performance of different pseudopotentials, $[Pb(H_2O)_n]^{2+}$ complexes were geometry-optimized using the Gaussian03 program[34] at the restricted Hartree−Fock (RHF) and B3LYP[35,36] levels of theory. Four different pseudopotential (PP)/basis set combinations were tested. Three large-core PPs were used: the Stuttgart relativistic large-core PP[37] (henceforth SDD), the large-core relativistic PP of Ross and co-workers[38] (henceforth CRENBS), and for consistency with previous SIBFA parametrization studies, the SBK compact relativistic PP.[39] Basis sets and PPs for Pb were taken from the Basis Set Exchange[40,41] and are included as Supporting Information. Remaining atom types were described using a 6-31+G** basis set for the SDD and CRENBS PPs, as successfully applied in previous work on $[Pb(H_2O)]^{2+}$.[42] SBK calculations employed an effective core potential for oxygen atoms[43] with a CEP 4−31G(2d) basis set for both oxygen and hydrogen. For comparison, a small-core relativistic PP of Peterson,[44] designed for used with aug-cc-pVnZ basis sets was selected and used with a triple-$\zeta$ basis set for lead and remaining atoms. Binding energies were evaluated as the difference in energy between the bound complex and isolated, geometry-optimized monomers.

Comparison of the performance of different ab initio and DFT methods was carried out similarly using the small-core PP of Peterson with aug-cc-pVTZ basis set for $Pb^{2+}$ and bound ligands. The DFT methods tested with this PP/basis set combination were the M05−2X,[45] M06,[46] B3LYP,[35,36,47] and PBE[48,49] density functionals available in Gaussian03.[34] Resticted Hartree−Fock, CCSD,[50,51] CCSD(T),[52] and MP2[53,54] calculations were performed with the same PP and aug-cc-pVTZ basis set. Finally, the $[Pb(H_2O)]^{2+}$ monohydrated complex was geometry-optimized with each method using a larger aug-ccpVQZ basis set and with the same small-core PP. This PP/quadruple-$\zeta$ basis set combination was additionally used with B3LYP to optimize $[Pb(H_2O)_n]^{2+}$ complexes ($n = 1,2,4$).

RVS calculations were performed using Gamess[55] to decompose the RHF interaction energy. The RVS procedure is related to Morokuma decomposition[56] but maintains the antisymmetry of the wave function, and in this respect is similar to the Constrained Space Orbital Variation (CSOV) method of Bagus.[57] PPs and basis sets used in Gamess were again taken from the Basis Set Exchange.

The Electron Localization Function (ELF), originally proposed by Becke and Edgecombe[58] offers useful insight

into the effect of complexation on the behavior of the $Pb^{2+}$ lone-pair. The basin associated with the lone pair V(Pb) is obtained by topological analysis to allow both quantitative study and visualization of the basin's position and shape.[58−64] In the present contribution, the ELF calculations and their analysis were performed using a modified version of the TOPMOD package[59,65,66] with electronic densities calculated by Gaussian03.

**SIBFA.** The SIBFA force field has been described in detail elsewhere,[15−17,19,24,67] only a brief overview will be given here with additional details available as Supporting Information. SIBFA is an anisotropic, polarizable force field in which molecules are broken down into rigid fragments that are free to rotate about their connection points (torsional degrees of freedom). Both inter- and intramolecular energies are evaluated as the sum of interactions between constituent chemical fragments. Unusually for a force field, SIBFA explicitly accounts for electrostatic ($E_{mtp}$), repulsion ($E_{rep}$), charge-transfer ($E_{ct}$), and polarization ($E_{pol}$) energies in evaluating the total interaction energy $\Delta E_{tot}$:

$$\Delta E_{tot} = E_{mtp} + E_{rep} + E_{pol} + E_{ct} + E_{disp} \qquad (1)$$

The final term "$E_{disp}$" represents an estimation of the dispersion energy to afford improved agreement with post-HF methodologies.[67]

The electrostatic term $E_{mtp}$ is evaluated using a multipole expansion, with both atom- and bond-centered multipole moments included up to quadrupole. Multipole moments are derived from the ab initio charge density according to the procedure of Vigné-Maeder and Claverie.[68] The sharing of multipoles between both atoms and bond centers ensures improved short-range convergence. A recently extended formulation[18] adds a penetration contribution to account for the overlap of molecular charge densities, further improving evaluation of short-range electrostatic interactions.

The short-range repulsion term $E_{rep}$ accounts for Pauli-repulsion between same-spin electrons, and is modeled in SIBFA by means of a sum of bond−bond, bond−lone pair and lone pair−lone pair interactions. Bond sites are located at the barycenters between bonded atoms within a fragment. The positions of electron lone pairs are defined using Boys localization of orbital centroids.[69−71] Use of bonds and lone-pairs in place of atom-centered repulsion sites allows increased anisotropy and a more accurate representation of the repulsion energy.[24,67]

The charge transfer term $E_{ct}$ is especially important where metal−ligand bonds exhibit some degree of covalent character. Electron density from the bound ligands can be transferred to the central metal cation *M*, and back-donation to or from the *d*-orbitals of transition metals can provide an additional energetic contribution.[17,72,73] Charge transfer in SIBFA is evaluated by approximating orbital overlap using the distance and angles between localized lone pair orbitals of the interacting entities.[17,72,74,75]

Polarization of the fragments' electron density makes the final major contribution to the total interaction energy in the standard SIBFA force field. The polarizability tensors are located on the centroids of the Boys localized orbitals (chemical bonds and heteroatom lone pairs). They are derived from the quantum chemical wave function of each fragment using a procedure developed by Garmer and Stevens.[69] Recent data show that, if off-centered lone pair polarizabilities are explicitly represented, classical polarizable force fields can afford a close agreement with the ab initio results, both in terms of polarization energy and in terms of dipole moment (see ref 71 for details). The polarization energy takes the form of an interaction between an induced dipole and the electric field generated by surrounding moieties; for metal cations an additional induced quadrupole interacting with the field gradient is considered (see references 19, 75 and references therein). A Gaussian screening term intervenes in the evaluation of the electric field arising from a given multipolar site at very close range to prevent close contact between charge-carrying sites of one moiety and polarizable sites of another, which could give rise to unphysically large polarization energies.[17,24]

## Results

**Pseudo-Potential Comparison.** A series of $[Pb(H_2O)_n]^{2+}$ clusters ($1 \leq n \leq 6$) were energy-minimized using the SDD, CRENBS, and SBK large-core PPs, as well as using one small-core PP with aug-cc-pVTZ basis set. Binding enthalpies of $Pb^{2+}$ were calculated as described in the Methods section. In the case of the hexacoordinated $[Pb(H_2O)_6]^{2+}$ complex two stable structures were found (Figure 1). The *holo*-directed $[Pb(H_2O)_6]^{2+}$ structure was found to be a stable minimum at both Hartree−Fock and B3LYP levels of theory (confirmed by frequency calculations) with all PPs. The *hemi*-directed structure was found to be a stable minimum using the Hartree−Fock method with all PPs, and with B3LYP when using all PPs except the small-core PP with aug-cc-pVTZ basis set. Additional MP2 optimizations with the small-core PP and aug-cc-pVTZ basis set did find a stable minimum, however, so the MP2 geometry of the *hemi*-directed structure was used to perform B3LYP single point calculations for comparison with the other methods. ELF isosurfaces are shown in Figure 1 to illustrate topological features of the sterically active $Pb^{2+}$ lone pair upon the transition from the *holo* to the *hemi*-directed structure of $[Pb(H_2O)_6]^{2+}$. Similar plots for $[Pb(CO)_5]^{2+}$ and $[Pb(CO)_6]^{2+}$ are provided as Figure S1 in Supporting Information and in previous work.[76]

Table 1 shows a comparison of the performance of the different PPs against previously published full-electron, four-component results for the $[Pb(H_2O)]^{2+}$ complex.[42] It can be seen that, while the small-core PP performs slightly better, SDD and CRENBS PPs also offer good agreement with the four-component results. SBK B3LYP results afford similar accuracy to CRENBS and SDD, although the binding energy is overestimated. The largest error of 6 kcal/mol arises with SBK and RHF. Pb−O distances are in good agreement for all PPs.

The final two columns of Table 1 show a comparison of binding energies using different PPs for the *holo*- and *hemi*-directed $[Pb(H_2O)_6]^{2+}$ clusters. There is now significant quantitative disagreement in the $Pb^{2+}$ binding energy, ranging, for example, from −207.7 kcal/mol for the *hemi*-directed structure with the SDD PP combined with B3LYP

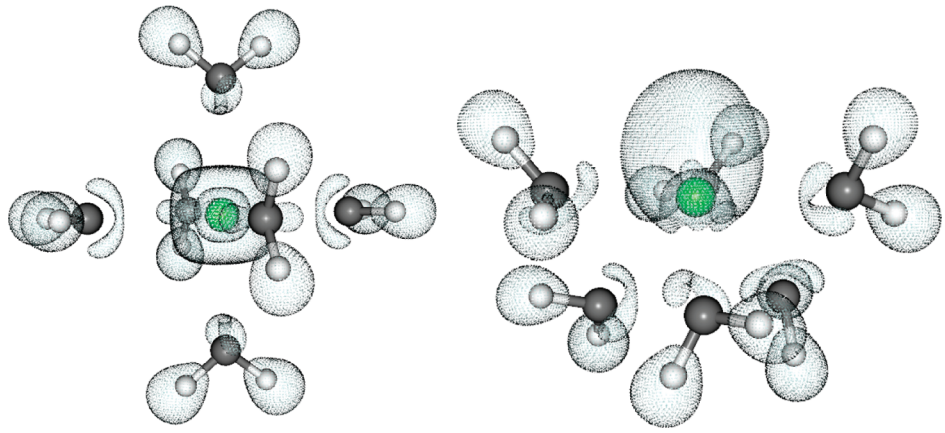


***Figure 1.*** ELF isosurfaces ($\eta = 0.085$) showing the $Pb^{2+}$ lone pair basin V(Pb) in *holo*- (left) and *hemi*-directed (right) $[Pb(H_2O)_6]^{2+}$ structures. $Pb^{2+}$ is shown in green with V(Pb) localized in the surrounding volume.

***Table 1.*** Comparison of Binding Energy and Pb−O Distance (Å) in the $[Pb(H_2O)]^{2+}$ Complex, and Binding Energies in *holo*- and *hemi*-Directed $[Pb(H_2O)_6]^{2+}$ Complexes Using Different Pseudopotentials

| | | $[Pb(H_2O)]^{2+}$ (BSSE corr) | | $(H_2O)_6$ *hemi* | $(H_2O)_6$ *holo* |
|---|---|---|---|---|---|
| pseudopotential | method | *r*(Pb−O) | Δ*E* (kcal/mol) | Δ*E* (kcal/mol) | Δ*E* (kcal/mol) |
| SBK | RHF | 2.320 | −59.5 | −238.1 | −234.5 |
| | B3LYP | 2.283 | −66.8 | −254.8 | −252.0 |
| CRENBS | RHF | 2.388 | −51.1 | −207.3 | −207.6 |
| | B3LYP | 2.367 | −56.1 | −218.5 | −219.3 |
| SDD | RHF | 2.384 | −49.9 | −197.6 | −197.8 |
| | B3LYP | 2.359 | −55.0 | −207.7 | −208.2 |
| aug-cc-pVTZ | RHF | 2.336 | −51.9 | −198.5 | −197.8 |
| | B3LYP | 2.322 | −58.5 | −211.6 | −215.0 |
| full-$e^-$ relativistic[a] | DHF | 2.347 | −53.5 | | |
| | DB3LYP | 2.338 | −61.0 | | |

[a] Values taken from earlier work of Gourlaouen et al.[42]

to −254.8 kcal/mol for the same complex with the SBK PP and B3LYP. Significantly, the predicted relative energies of the *holo*- and *hemi*-directed structures also differ. While the small-core PP suggests that the *holo*-directed structure is around 3.4 kcal/mol more stable than the *hemi*-directed structure, SDD and CRENBS with B3LYP predict the *holo*-directed structure to be approximately iso-energetic with the *hemi*-directed conformation. In contrast, SBK predicts the *hemi*-directed structure to be 2.8 kcal/mol more stable than the *holo* arrangement. The PP chosen therefore has a non-negligible impact on both the magnitude of the binding energy and the relative energies of different *holo*- and *hemi*-directed conformations of the same complex, necessitating careful selection.

Final selection of an appropriate PP was made on the basis of RVS data. Figure 2 shows the RVS Coulomb and repulsion energies in the $[Pb(H_2O)]^{2+}$ complex as a function of Pb−O distance. While the Coulomb energy becomes exponentially more attractive with decreasing ligand-cation distance using the small-core PP, as should intuitively be the case, the large-core PPs cause artifacts to arise at short distance as the electron density of the ligand begins to overlap the effective core potential. The divergence between data from large and small core PPs starts to occur at around 2.4 Å, the optimized bond length in many of the Pb−$H_2O$ complexes. The repulsion energy is also significantly lower at short-range using the large-core PPs than using the small-core PP. This leads to some cancellation of errors with the correspondingly lower Coulomb contribution, but divergence in the repulsion energy starts at longer range than divergence in the Coulomb energy (about 2.6 Å). As the small-core PP affords the best agreement with four-component calculations for the $[Pb(H_2O)]^{2+}$ complex and appears free from artifacts arising from ligand-PP core overlap at the distances of interest, it was selected to provide reference data for the series of $[Pb(H_2O)_n]^{2+}$ complexes.

**Ab Initio/Density Functional Method Comparison.** Having selected the small-core PP with aug-cc-pVTZ basis set, the stability of data to changes in electronic structure method was next investigated. Results are presented in Table 2. The less computationally expensive DFT methodologies were applied to all systems up to $[Pb(H_2O)_5]^{2+}$, whereas CCSD optimizations were only possible up to $[Pb(H_2O)_2]^{2+}$, and CCSD(T) was only possible for the $[Pb(H_2O)]^{2+}$ complex with this PP/basis set combination. The table shows that all methods yield similar results. In particular, all DFT methods agree closely with one another. They afford complex formation energies around 3 kcal/mol lower than CCSD and CCSD(T) results for the $[Pb(H_2O)]^{2+}$ complex, with MP2 results lying between the two. Hartree−Fock binding energies are somewhat underestimated with respect to the other methods. The close agreement between DFT and MP2 results is also visible for the larger clusters, suggesting that choice of electronic structure method has a smaller impact on calculated binding energies than the choice of PP.

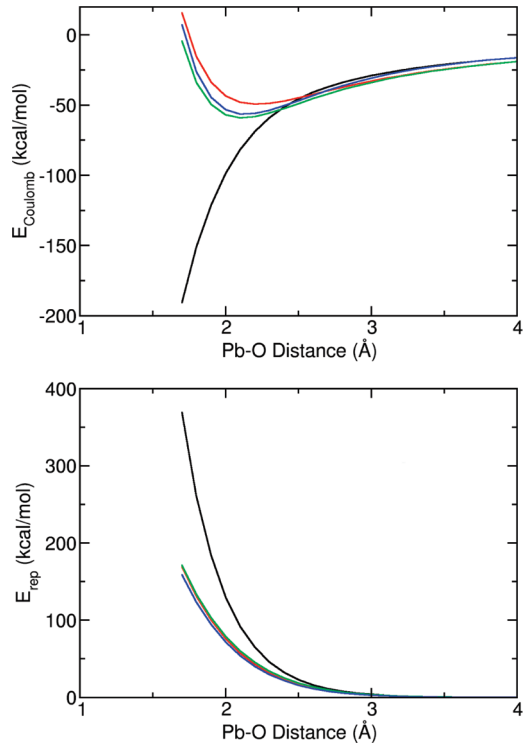



***Figure 2.*** Coulomb (top) and repulsion (bottom) RVS contributions to the total interaction energy using different pseudopotentials: small-core with aug-cc-pVTZ (black), SDD (red), CRENBS (green), and SBK (blue) pseudopotentials with different basis sets are compared (see text for details).

***Table 2.*** Comparison of $[Pb(H_2O)_n]^{2+}$ Complex Formation Energies Relative to Gas Phase Monomer Energies Using the aug-cc-pVTZ Small-Core Pseudopotential and Basis Set with Different ab Initio and Density Functional Methods[a]

| $[Pb(H_2O)]^{2+}$ | $[Pb(H_2O)_2]^{2+}$ | $[Pb(H_2O)_4]^{2+}$ | $[Pb(H_2O)_5]^{2+}$ | |
|---|---|---|---|---|
| RHF | −51.9 | −95.0 | −157.8 | −180.4 |
| B3LYP pVTZ | −58.5 | −105.5 | −171.6 | −195.4 |
| B3LYP pVQZ | −58.8 | −106.0 | −172.2 | – |
| M05−2X | −59.3 | −108.4 | −180.5 | −206.9 |
| M06 | −58.6 | −106.9 | −177.9 | −204.0 |
| PBE | −59.3 | −107.4 | −175.8 | −200.1 |
| MP2 | −56.6 | −103.5 | −172.4 | −197.9 |
| CCSD | −55.3 | −100.6 | | |
| CCSD(T) | −55.7 | | | |
| full $e^-$ DB3LYP[b] | −61.0 | | | |

[a] The same pseudopotential with a larger aug-cc-pVQZ basis set is also shown. [b] Values taken from earlier work of Gourlaouen et al.[40]

Finally, the convergence of the basis set was examined further by running a limited number of optimizations using the larger aug-cc-pVQZ basis set with B3LYP and the small-core PP. Again, little change in binding energies is observed, with differences between the aug-cc-pVTZ and aug-cc-pVQZ results generally within 1 kcal/mol. Similar results (not shown) were obtained using this basis set and the other electronic structure methods for the monohydrated complex. The B3LYP method with aug-cc-pVTZ basis set therefore appears to represent an acceptable level of theory to provide reference data for subsequent force field parametrization. In addition to good agreement with other methods tested here, it has been widely applied to the many organic ligands that will be of interest when applying the force field to studies of biological systems.

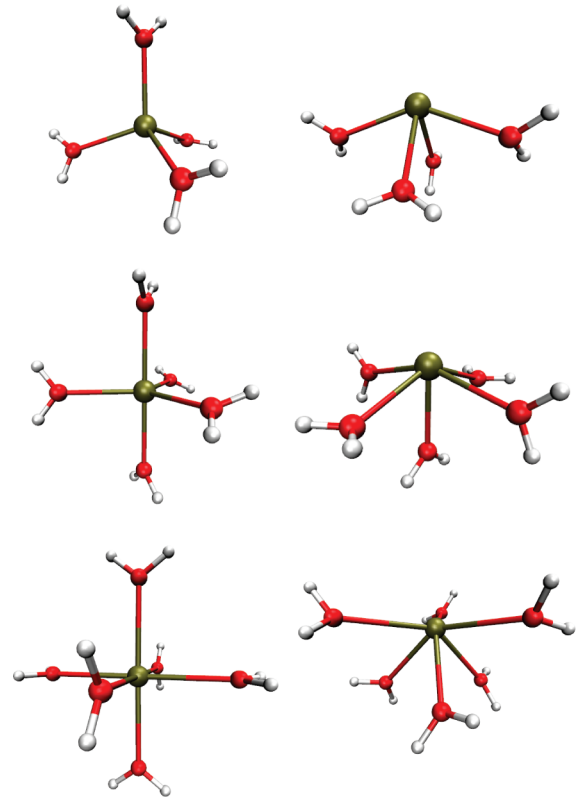


***Figure 3.*** $[Pb(H_2O)_4]^{2+}$ (top), $[Pb(H_2O)_5]^{2+}$ (middle) and $[Pb(H_2O)_6]^{2+}$ (bottom) *holo*- and *hemi*-directed structures used to investigate the energetic origins of *holo*- and *hemi*-directionality.

**RVS Analysis of *holo*- and *hemi*-Directed Structures.** The energetic origins of the stabilization of *hemi*-directed complexes over their *holo*-directed counterparts were next investigated using RVS analysis. A series of $[Pb(H_2O)_4]^{2+}$, $[Pb(H_2O)_5]^{2+}$, and $[Pb(H_2O)_6]^{2+}$ complexes were selected for study (Figure 3). The energy-minimized *hemi*-directed structures of all complexes obtained with the small-core PP and aug-cc-pVTZ basis set and B3LYP method were used, although, as already stated, an MP2 geometry had to be taken for $[Pb(H_2O)_6]^{2+}$ as no stable *hemi*-directed structure was found for this complex using a small-core PP with B3LYP. The *holo*-directed structure of $[Pb(H_2O)_6]^{2+}$ was used for comparison with the *hemi*-directed structure, along with an artificially created tetrahedral $[Pb(H_2O)_4]^{2+}$ complex and a trigonal-bipyramidal $[Pb(H_2O)_5]^{2+}$ structure that were obtained by constrained optimization of Pb-ligand bond lengths. These artificial *holo*-directed structures were created to allow direct RVS comparison for each complex between a *holo*- and the corresponding *hemi*-directed ligand arrangement.

The results of the RVS analysis are shown in Table 3. Interestingly, in all three complexes the lower ligand−ligand repulsion energy associated with the *holo*-directed complexes more than compensates for the slightly lower Coulomb energy, meaning that the total first-order con-

***Table 3.*** RVS Energy Decomposition of *holo*- and *hemi*-Directed Structures of $[Pb(H_2O)_4]^{2+}$, $[Pb(H_2O)_5]^{2+}$, and $[Pb(H_2O)_6]^{2+}$ Complexes (kcal/mol)

| | $[Pb(H_2O)_4]^{2+}$ | | $[Pb(H_2O)_5]^{2+}$ | | $[Pb(H_2O)_6]^{2+}$ | |
|---|---|---|---|---|---|---|
| $[Pb(H_2O)_4]^{2+}$ | *hemi* | *holo* | *hemi* | *holo* | *hemi* | *holo* |
| Coulomb | −179.7 | −171.4 | −206.4 | −198.8 | −229.0 | −224.3 |
| repulsion | 103.5 | 87.5 | 106.1 | 91.9 | 108.2 | 94.5 |
| polarization | −70.2 | −61.2 | −71.2 | −63.9 | −70.4 | −64.6 |
| Pb-polarization | −7.4 | −1.6 | −6.1 | −1.2 | −6.0 | −0.6 |
| charge-transfer | −18.3 | −15.8 | −18.5 | −16.3 | −18.2 | −16.7 |
| total interaction | −159.7 | −153.8 | −182.5 | −177.7 | −200.1 | −199.4 |

tribution favors *holo*-directionality. This is also the case in the tetrahedral and trigonal-bipyramidal $[Pb(H_2O)_4]^{2+}$ and $[Pb(H_2O)_5]^{2+}$ complexes. Second-order energetic terms (polarization and charge-transfer) therefore account for the stabilization of *hemi*-directed structures in this series of complexes. The polarization energy makes the largest second order stabilizing contribution, with most of the difference between *holo*- and *hemi*-directed conformations arising from polarization of the $Pb^{2+}$ cation. This result can be rationalized, as arranging ligands on one side of the $Pb^{2+}$ cation generates a net electric field at the position of the metal cation whereas field-cancellation arises almost completely from evenly spacing ligands in opposing positions in a *holo*-oriented complex. Indeed, the difference in cation polarization energy between the $[Pb(H_2O)_4]^{2+}$ and $[Pb(H_2O)_5]^{2+}$ *holo*- and *hemi*-directed complexes is roughly equal to the difference in total binding energy favoring the *hemi*-directed structure. Charge-transfer, too, is non-negligible, and adds a smaller contribution of around 2 kcal/mol favoring *hemi*-directionality. As a relatively large basis set is used, BSSE distortion of the charge transfer energy is small (of the order of 0.01 kcal/mol) so the estimates are considered to be reliable. It should be noted that RVS total interaction energies are expected to differ slightly to the Gaussian03 binding energies reported above, as the complexation energy in RVS calculations is calculated as the difference between the isolated monomers in their supermolecular conformations, rather than in their gas phase-optimized geometries as reported in Tables 1 and 2.

**Parametrization of SIBFA.** An accurate representation of cation polarization and charge transfer, then, is key to modeling the energies of $Pb^{2+}$ complexes. While such terms are missing from simpler force fields, detailed approaches such as SIBFA are better equipped for such tasks. The adjustable parameters required in SIBFA to describe the separate energetic contributions of eq 1 were therefore fitted using RVS results for the $[Pb(H_2O)]^{2+}$ complex as a function of Pb−O distance. The results of this fitting are shown graphically in Figure 4 and are tabulated in Table 4. A good fit was possible for all terms, although the polarization energy deviates slightly from the corresponding RVS value at short-range. The cation quadrupolar polarization energy terms were fitted similarly using a $[Pb(H_2O)_2]^{2+}$ complex with $H_2O$ ligands placed opposite one another to neutralize the field generated at the $Pb^{2+}$ position, leaving a field gradient. An acceptable fit was again achieved as a function of the Pb−O bond length.

Validation of the SIBFA parameters was performed by modeling the six complexes of *holo*- and *hemi*-directed $[Pb(H_2O)_4]^{2+}$, $[Pb(H_2O)_5]^{2+}$, and $[Pb(H_2O)_6]^{2+}$ presented in Figure 3. As shown in Table 5, very satisfactory agreement is observed between SIBFA and RVS results. All contributions are well represented, including the

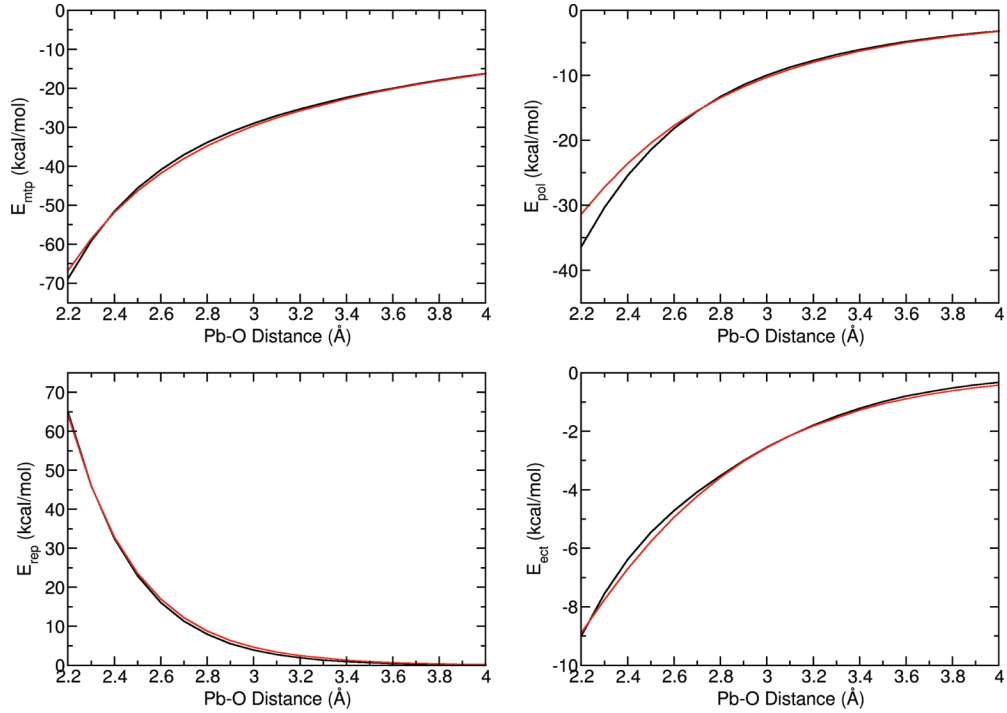



***Figure 4.*** RVS (black) vs SIBFA (red) Coulomb (top left), polarization (top right), repulsion (bottom left) and charge-transfer (bottom right) energies as a function of the Pb−O distance in $[Pb(H_2O)]^{2+}$.

***Table 4.*** SIBFA vs RVS Data at Different Pb−O Separations *r* in the $[Pb(H_2O)]^{2+}$ Complex[a]

| | *E*(mtp) | | *E*(rep) | | *E*(pol) | | *E*(ct) | |
|---|---|---|---|---|---|---|---|---|
| *r* (Å) | RVS | SIBFA | RVS | SIBFA | RVS | SIBFA | RVS | SIBFA |
| 2.2 | −68.89 | −66.88 | 65.08 | 64.25 | −36.46 | −31.43 | −9.01 | −8.90 |
| 2.3 | −59.11 | −58.54 | 45.98 | 45.87 | −30.34 | −27.20 | −7.55 | −7.75 |
| 2.4 | −51.55 | −51.82 | 32.44 | 32.84 | −25.43 | −23.56 | −6.38 | −6.71 |
| 2.5 | −45.63 | −46.34 | 22.87 | 23.57 | −21.45 | −20.44 | −5.46 | −5.77 |
| 2.6 | −40.91 | −41.81 | 16.09 | 16.95 | −18.18 | −17.76 | −4.71 | −4.94 |
| 2.7 | −37.09 | −38.02 | 11.31 | 12.22 | −15.51 | −15.47 | −4.08 | −4.21 |
| 2.8 | −33.94 | −34.81 | 7.94 | 8.82 | −13.31 | −13.50 | −3.52 | −3.58 |
| 2.9 | −31.29 | −32.06 | 5.57 | 6.38 | −11.50 | −11.81 | −3.01 | −3.03 |
| 3.0 | −29.02 | −29.69 | 3.90 | 4.62 | −10.01 | −10.36 | −2.55 | −2.56 |
| 3.1 | −27.05 | −27.61 | 2.73 | 3.35 | −8.77 | −9.11 | −2.15 | −2.16 |
| 3.2 | −25.32 | −25.78 | 1.91 | 2.44 | −7.73 | −8.04 | −1.80 | −1.81 |
| 3.3 | −23.78 | −24.23 | 1.33 | 1.87 | −6.83 | −7.18 | −1.48 | −1.55 |
| 3.4 | −22.39 | −22.70 | 0.93 | 1.29 | −6.08 | −6.31 | −1.22 | −1.27 |
| 3.5 | −21.14 | −21.38 | 0.65 | 0.94 | −5.43 | −5.62 | −0.99 | −1.07 |
| 3.6 | −20.00 | −20.19 | 0.45 | 0.69 | −4.85 | −5.02 | −0.80 | −0.89 |
| 3.7 | −18.95 | −19.11 | 0.31 | 0.50 | −4.36 | −4.50 | −0.65 | −0.74 |
| 3.8 | −17.99 | −18.12 | 0.22 | 0.37 | −3.93 | −4.04 | −0.52 | −0.62 |
| 3.9 | −17.10 | −17.21 | 0.15 | 0.27 | −3.54 | −3.64 | −0.42 | −0.51 |
| 4.0 | −16.28 | −16.36 | 0.11 | 0.20 | −3.20 | −3.28 | −0.33 | −0.43 |

[a] Coulomb ($E_{mtp}$), repulsion ($E_{rep}$), polarization ($E_{pol}$), and charge-transfer ($E_{ct}$) energies (kcal/mol) are compared at each separation.

***Table 5.*** Comparison of RVS Energy Decomposition and Corresponding SIBFA Energetic Components for $[Pb(H_2O)_n]^{2+}$ Complexes in *holo*- and *hemi*-Directed Conformations (kcal/mol)[a]

| | $[Pb(H_2O)_4]^{2+}$ | | | | $[Pb(H_2O)_5]^{2+}$ | | | | $[Pb(H_2O)_6]^{2+}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RVS | | SIBFA | | RVS | | SIBFA | | RVS | | SIBFA | |
| | *hemi* | *holo* | *hemi* | *holo* | *hemi* | *holo* | *hemi* | *holo* | *hemi* | *holo* | *hemi* | *holo* |
| Coulomb ($E_{mtp}$) | −179.7 | −171.4 | −181.1 | −172.8 | −206.4 | −198.8 | −210.0 | −201.9 | −229.0 | −224.3 | −235.9 | −229.2 |
| repulsion ($E_{rep}$) | 103.5 | 87.5 | 103.1 | 85.1 | 106.1 | 91.9 | 106.9 | 91.1 | 108.2 | 94.5 | 111.8 | 95.3 |
| pol. RVS ($E_{pol}$*) | −70.2 | −61.2 | −64.9 | −59.7 | −71.2 | −63.9 | −67.6 | −63.2 | −70.4 | −64.6 | −67.7 | −64.8 |
| pol. VL ($E_{pol}$) | −65.3 | −54.2 | −67.6 | −55.0 | −64.3 | −54.5 | −66.1 | −56.1 | −61.1 | −53.1 | −63.3 | −55.9 |
| Pb-polarization | −7.4 | −1.6 | −5.3 | 0.0 | −6.1 | −1.2 | −4.5 | −0.1 | −6.0 | −0.6 | −4.7 | 0.0 |
| charge-transfer ($E_{ct}$) | −18.3 | −15.8 | −17.0 | −16.4 | −18.5 | −16.3 | −18.1 | −17.5 | −18.2 | −16.7 | −19.0 | −18.2 |
| total | −159.7 | −153.8 | −162.7 | −159.1 | −182.5 | −177.7 | −187.3 | −184.5 | −200.1 | −199.4 | −206.4 | −207.9 |

[a] The polarization energy is shown before iteration (Pol. RVS/$E_{pol}$*) and after iteration (Pol. VL/$E_{pol}$) to self-consistency of the electric field (see text for details). In column 1, headings describe the RVS contributions to the total interaction energy, with values in parentheses indicating which term in SIBFA this corresponds to (see eq 1).

polarization energy of each system before and after iteration of the complex's electric field to self-consistency. The "RVS" polarization energy is the same polarization energy reported in Tables 3 and 4 and arises from the polarization of each monomer due to the unperturbed electric field created by each surrounding *unpolarized* monomer. The 'variation-like' (VL) value arises from polarization of each monomer by the electric field generated by each of the surrounding *polarized* monomers. This second value requires an iterative, self-consistent procedure in SIBFA. Not only are absolute values of the different contributions of the interaction energy in good quantitative agreement with RVS results, but importantly the relative energies of the different *holo*- and *hemi*-directed structures are also in good agreement. The *hemi*-directed structures are successfully predicted to be more stable in both the $[Pb(H_2O)_4]^{2+}$ and $[Pb(H_2O)_5]^{2+}$ complexes, while the iso-energetic $[Pb(H_2O)_6]^{2+}$ *holo*- and *hemi*-directed structures are predicted to be separated by 1.5 kcal/mol only. SIBFA is therefore able to encapsulate the energetic processes underlying stabilization of *hemi*-directed structures, at least for the Pb−$H_2O$ complexes presented, by means of an accurate treatment of cation polarization.

As a final point of interest, QM/MM calculations were used to investigate whether point charge representations of the $H_2O$ ligands, of the type widely applied in classical force fields, were sufficient to induce a characteristic *hemi*-directed shift in lone-pair distribution such as that visible in Figure 1. These QM/MM calculations were performed with Gaussian03 at the B3LYP level of theory. $Pb^{2+}$ was again represented using the small-core pseudopotential and aug-cc-pVTZ basis-set, while Kollman[77] charges fitted to B3LYP/aug-cc-pVTZ data were used to represent ligand atoms. As shown in Figure 5, a *hemi*-directed arrangement of point charges representing the $H_2O$ ligands is insufficient to significantly displace the $Pb^{2+}$ lone pair basin in the $[Pb(H_2O)_5]^{2+}$ complex. The ELF analysis of the QM-represented cation therefore shows an essentially spherical lone-pair distribution, with the cation nucleus located roughly at the center. This result demonstrates that, while an accurate treatment of cation polarization is vital to describe Pb-ligand interactions, a realistic representation

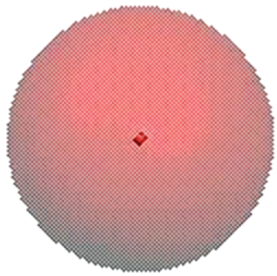


***Figure 5.*** ELF isosurface ($\eta = 0.8$) showing the V(Pb) basin in QM/MM representation of *hemi*-directed $[Pb(H_2O)_5]^{2+}$ structure: the use of classical force field point charges to represent $H_2O$ ligands results in only a very slight distortion of the QM $Pb^{2+}$ V(Pb) basin from the spherical distribution observed for the *holo*-directed structure.

of ligand electrostatics and other energetic contributions must also be maintained.

## Conclusions

A series of hydrated $Pb^{2+}$ complexes have been used to study the underlying physical origins of the stabilization of *holo*- or *hemi*-directed arrangements of ligands from an energy decomposition perspective. The RVS analysis demonstrates that the stabilization of *hemi*-directed structures can be explained largely in terms of cation polarization arising from the electric field (and to a lesser extent from the field gradient) generated by the hemitropic ligand arrangement. Corresponding *holo*-directed structures, while reducing ligand−ligand repulsion, lead to small or zero net electric fields and, consequently, lower polarization energy.

Parameters fitted for the SIBFA force field have demonstrated that an accurate representation of the cation polarization can encapsulate the *hemi*-stabilizing effect. Good quantitative agreement was achieved in total complex binding energies and in relative *holo*- and *hemi*-directed complex energies in the series of hydrated Pb clusters studied. Extension of the force field to additional ligand types is currently underway, allowing molecular mechanics investigation of the interactions of $Pb^{2+}$ cations within complexes[11,76] and biological systems[78] and aiding in the search for selective chelating agents that can be used in vivo to cure lead poisoning.

Finally, it is found that the level of theory employed, in particular the choice between large-core and small-core PPs, can have a significant impact on calculated binding energies and especially on RVS values at short ligand-$Pb^{2+}$ distances. RVS results suggest that this discrepancy arises primarily from artifacts associated with the overlap of ligand electronic density and that arising from the large-core PPs at short-range. Small-core PPs therefore represent a safer choice for studies of this kind.

**Acknowledgment.** The computations reported in this work were performed using resources from GENCI (CINES/IDRIS), Grant Nos. 2009-075009 and x2010075027, and from the Centre de Ressources Informatiques de Haute Normandie (CRIHAN, Rouen, France), Grant Nos. 1998053 and 2008011. Financial support from the French National Research Agency (ANR) on project SATURNIX (No. 2008-CESA-020) is acknowledged.

**Supporting Information Available:** The pseudopotentials used in this work. Also included is a figure demonstrating ELF analysis of $[Pb(CO)_5]^{2+}$ and $[Pb(CO)_6]^{2+}$ complexes, and additional details relating to the SIBFA force field. This information is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Patterson, C. C. *Arch. Environ. Health* **1965**, *11*, 344–360.

(2) Casas, J.; Fernández, J. S. C.; Sordo, J. *Lead: Chemistry, Analytical Aspects, Environmental Impact and Health Effects*; Elsevier: New York, 2006.

(3) Finkelstein, Y.; Markowitz, M. E.; Rosen, J. F. *Brain Res. Rev.* **1998**, *27*, 168–176.

(4) Wang, S.; Zhang, J. *J. Environ. Res.* **2006**, *101*, 412–418.

(5) Bergdahl, I. A.; Grubb, A.; Schütz, A.; Desnick, R. J.; Wetmur, J. G.; Sassa, S.; Skerfving, S. *Pharmacol. Toxicol.* **1997**, *81*, 153–158.

(6) Gourlaouen, C.; Parisel, O. *Ang. Chem. Intern.* **2007**, *46*, 553–556.

(7) Ghering, A. B.; Miller Jenkins, L. M.; Schenck, B. L.; Deo, S.; Mayer, R. A.; Pikaart, M. J.; Omichinski, J. G.; Godwin, H. A. *J. Am. Chem. Soc.* **2005**, *127*, 3751–3759.

(8) Godwin, H. A. *Curr. Op. Chem. Bio.* **2001**, *5*, 223–227.

(9) Jaffe, E. K.; Martins, J.; Li, J.; Kervinen, J.; Dunbrack, R. L. *J. Biol. Chem.* **2001**, *276*, 1531–1537.

(10) Gourlaouen, C.; Parisel, O. *Int. J. Quantum Chem.* **2008**, *108*, 1888–1897.

(11) van Severen, M.-C.; Gourlaouen, C.; Parisel, O. *J. Comput. Chem.* **2010**, *31*, 185–194.

(12) Shimoni-Livny, L.; Glusker, J. P.; Bock, C. W. *Inorg. Chem.* **1998**, *37*, 1853–1867.

(13) van Severen, M.-C.; Piquemal, J.-P.; Parisel, O. *Chem. Phys. Lett.* **2009**, *478*, 17–19.

(14) Stevens, W. J.; Fink, W. *Chem. Phys. Lett.* **1987**, *139*, 15–22.

(15) Piquemal, J.-P.; Williams-Hubbard, B.; Fey, N.; Deeth, R. J.; Gresh, N.; Giessner-Prettre, C. *J. Comput. Chem.* **2003**, *24*, 1963–1970.

(16) Gresh, N.; Claverie, P.; Pullman, A. *Theor. Chim. Acta* **1984**, *66*, 1–20.

(17) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856–882.

(18) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353–10359.

(19) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.

(20) Gresh, N.; Stevens, W. J.; Krauss, M. *J. Comput. Chem.* **1995**, *16*, 843–855.

(21) Gresh, N. *J. Phys. Chem. A* **1997**, *101*, 8680–8694.

(22) de Courcy, B.; Pedersen, L. G.; Parisel, O.; Gresh, N.; Silvi, B.; Pilmé, J.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2010**, *6*, 1048–1063.

(23) (a) Antony, J.; Piquemal, J. P.; Gresh, N. *J. Comput. Chem.* **2005**, *26*, 1131–1147. (b) Gresh, N.; Piquemal, J.-P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113–1130.

(24) (a) Roux, C.; Gresh, N.; Perera, L.; Piquemal, J.-P.; Salmon, L. *J. Comput. Chem.* **2007**, *28*, 938–957. (b) Jenkins, L. M. M.; Hara, T.; Durell, S. R.; Hayashi, R.; Inman, J. K.; Piquemal, J.-P.; Gresh, N.; Appella, E. *J. Am. Chem. Soc.* **2007**, *129*, 11067–11078. (c) Foret, F.; de Courcy, B.; Gresh, N.; Piquemal, J.-P.; Salmon, L. *Bioorg. Med. Chem.* **2009**, *17*, 7100–7107. (d) Gresh, N.; Audiffren, N.; Piquemal, J.-P.; de Ruyck, J.; Ledecq, M.; Wouters, J. *J. J. Phys. Chem. B* **2010**, *114*, 4884–4895.

(25) de Courcy, B.; Piquemal, J.-P.; Garbay, C.; Gresh, N. *J. Am. Chem. Soc.* **2010**, *132*, 3312–3320.

(26) de Courcy, B.; Piquemal, J.-P.; Gresh, N. *J. Chem. Theory Comput.* **2008**, *4*, 1659–1668.

(27) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, I. W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(28) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(29) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(30) Åstrand, P.-O.; Linse, P.; Karlström, G. *J. Chem. Phys.* **1995**, *191*, 195–202.

(31) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6365–6376.

(32) Schaeffer, C. E.; Jorgensen, C. K. *Mol. Phys.* **1965**, *9*, 401–412.

(33) Larsen, E.; Mar, G. N. L. *J. Chem. Educ.* **1974**, *51*, 633–640.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision E.01*; Gaussian, Inc., Wallingford, CT, 2004.

(35) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

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

(37) Küchle, W.; Dolg, M.; Stoll, H.; Preuss, H. *Mol. Phys.* **1991**, *74*, 1245–1263.

(38) Ross, R. B.; Powers, J. M.; Atashroo, T.; Ermler, W. C.; LaJohn, L. A.; Christiansen, P. A. *J. Chem. Phys.* **1990**, *93*, 6654–6670.

(39) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612–630.

(40) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

(41) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571–1586.

(42) Gourlaouen, C.; Piquemal, J.-P.; Parisel, O. *J. Chem. Phys.* **2006**, *124*, 174311.

(43) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026–6033.

(44) Peterson, K. A. *J. Chem. Phys.* **2003**, *119*, 11099–11112.

(45) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(46) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *120*, 215–241.

(47) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(48) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(49) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(50) Purvis, G. D. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(51) Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F. *J. Chem. Phys.* **1988**, *89*, 7382–7387.

(52) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

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

(54) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(55) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(56) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325–340.

(57) Bagus, P. S.; Hermann, K.; Bauschlicher, C. W. *J. Chem. Phys.* **1984**, *80*, 4378–4386.

(58) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–5403.

(59) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem. (Oxford)* **1999**, *23*, 597–604.

(60) Savin, A.; Nesper, R.; Wengert, S.; Fässler, T. F. *Angew. Chem.* **1997**, *109*, 1892–1918.

(61) Savin, A.; Nesper, R.; Wengert, S.; Fässler, T. F. *Ang. Chem. Intern.* **1997**, *36*, 1808–1832.

(62) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683–686.

(63) Savin, A.; Silvi, B.; Coionna, F. *Can. J. Chem.* **1996**, *74*, 1088–1096.

(64) Piquemal, J.-P.; Pilmé, J.; Parisel, O.; Gérard, H.; Fourré, I.; Bergès, J.; Gourlaouen, C.; De la Lande, A.; van Severen, M.-C.; Silvi, B. *Int. J. Quantum Chem.* **2008**, *108*, 1951–1969.

(65) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. TopMod Package. This package is available on the web site of the Laboratoire

de Chimie Théorique, Université Pierre et Marie Curie (UMR 7616, CNRS—Paris 6 -UPMC), URL: www.lct.jussieu.fr (see the personal home page of Prof. B. Silvi) Accessed July 2010.; 1997.

(66) Pilmé, J.; Piquemal, J.-P. *J. Comput. Chem.* **2008**, *29*, 1440–1449.

(67) Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824–837.

(68) Vigne-Maeder, F.; Claverie, P. *J. Chem. Phys.* **1988**, *88*, 4934–4948.

(69) Garmer, D. R.; Stevens, W. J. *J. Phys. Chem.* **1989**, *93*, 8263–8270.

(70) (a) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300–302. (b) Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 296–299.

(71) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, *111*, 8170–8176.

(72) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1986**, *29*, 101–118.

(73) Murrell, J. N.; Randic, M.; Williams, D. R. *Proc. R. Soc. London* **1965**, *284*, 566–581.

(74) (a) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1985**, *28*, 757–771. (b) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1982**, *22*, 199–215.

(75) Gresh, N.; Policar, C.; Giessner-Prettre, C. *J. Phys. Chem. A* **2002**, *106*, 5660–5670.

(76) Gourlaouen, C.; Gérard, H.; Piquemal, J.-P.; Parisel, O. *Chem.—Eur. J.* **2008**, *14*, 2730–2743.

(77) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

(78) van Severen, M.-C.; Piquemal, J.-P.; Parisel, O. *J. Phys. Chem. B* **2010**, *114*, 4005–4009.

CT1004005

JCTC Journal of Chemical Theory and Computation

# Accurate Vibrational Frequencies of Borane and Its Isotopologues

Patrick Meier, Michael Neff, and Guntram Rauhut*

*Institut für Theoretische Chemie, Universität Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany*

Received August 24, 2010

**Abstract:** Vibrational transitions of borane and its isotopologues ($^{11}BH_3$, $^{11}BD_3$, $^{10}BH_3$, and $^{10}BD_3$) have been obtained from state-specific vibrational configuration interaction calculations. Explicitly correlated coupled-cluster calculations, CCSD(T)-F12a, with additional corrections for high-order terms of the coupled-cluster expansion, i.e., CCSDT(Q), were used to determine multidimensional potential energy surfaces. Additional contributions due to core−valence interactions, scalar relativistic effects, and such arising from the diagonal Born−Oppenheimer correction were accounted for in the one-dimensional terms within the expansion of the potential energy surface. From these, anharmonic vibrational spectra were obtained, which are in excellent agreement with experimental data. Mean absolute deviations from gas phase measurements were found to be in the sub-wavenumber regime.

## I. Introduction

The vibrational spectrum of borane has been the subject of many theoretical and experimental studies.[1−9] The empty p orbital accounts for the high reactivity of borane and thus explains its rapid dimerization to the more stable diborane. As a consequence, the measurement of borane by means of spectroscopic techniques is a tedious task, and many attempts failed until Kaldor and Porter[1] and later on Kawaguchi et al.[2−4] successfully reported the detection of the fundamental vibrational transitions. For a discussion of the *borane story*, see the papers of Galbraith et al.[7] and Martin and Lee.[5] From the computational point of view, borane is a challenging system, because its low number of electrons allows for calculations at the highest level, and thus it has been in the focus of theoreticians for a long time. Botschwina,[6] Galbraith et al.,[7] Martin and Lee,[5] and Feller et al.[8] studied the harmonic spectrum of $BH_3$ and its thermochemistry at different levels of electronic structure theory. From these, Botschwina[6] estimated the anharmonic transitions by scaling and thus provided the first reliable predictions until Martin and Lee used a quartic force field based on CCSD(T) calculations in combination with vibrational perturbation theory for a direct and fully *ab initio* calculation of the fundamentals.[5] A few years later, Schwenke[9] used the quartic force field of Martin and Lee to perform variational calculations. As a result, many spectroscopic constants are available for the $^{11}BH_3$ and $^{10}BH_3$ isotopologues, while theoretical studies focusing on anharmonic frequencies are still missing for the deuterated species. Moreover, vibrational overtones and combination bands have not yet been calculated, except for a very few. The motivation of this study is two-fold: First, we intend to provide very accurate anharmonic transitions for those isotopologues for which experimental data are not yet available, while verifying our computational scheme on the basis of those fundamentals, which have accurately been determined by experiment. Second, we would like to show that fully automated *ab initio* approaches are capable of predicting vibrational frequencies with very high accuracy.

Our approach for calculating the anharmonic transitions of the borane isotopologues differs substantially from the approach of Schwenke,[9] although it is also based on the variational principle. We use the Watson Hamiltonian and thus expand the potential in terms of normal coordinates, rather than symmetry adapted internal coordinates.

$$V(\mathbf{q}) = \sum_{ri} V_{ri}\left[p_r^{(i)} + \frac{1}{2}\sum_{sj} V_{sj}\left[p_{rs}^{(ij)} + \frac{1}{3}\sum_{tk} V_{tk}[p_{rst}^{(ijk)} + ...]\right]\right] \quad (1)$$

In this expression, $V_{ri}$ denotes single particle potentials, and $p_r^{(i)}$ denotes the corresponding coefficients of the expansion. The

* To whom correspondence should be addressed. E-mail: rauhut@theochem.uni-stuttgart.de.

10.1021/ct1004752 
Published on Web 12/08/2010

potential is calculated in a fully automated manner, which allows us to use any electronic structure level as implemented in the Molpro suite of *ab initio* programs.[10] One mode wave functions (modals) are obtained from state-specific vibrational self-consistent field calculations (VSCF)[11−13] and a subsequent accounting of correlation effects by means of vibration configuration interaction calculations (VCI).[14]

## II. Computational Details

**A. Electronic Structure Calculations.** Geometries, normal coordinates, and potential energy surfaces were obtained from explicitly correlated coupled-cluster theory, CCSD(T)-F12a, in combination with a triple-ζ basis set, i.e., vtz-f12.[15] As we have shown recently, results obtained at this level are extremely close to the basis set limit, and thus we consider effects due to the incompleteness of the orbital basis to be essentially negligible.[16] Slater geminals with an exponent of 1.0 $a_0^{-1}$ were used throughout. For the resolution of identity approximation (RI), a complementary auxiliary basis set (CABS) was built from the cc-pVTZ/JKFIT basis and the atomic orbital basis.[17] An aug-cc-pVTZ/MP2FIT basis was used within the density fitting of the integrals.[18] The CABS singles correction to the Hartree−Fock energies was included in all F12 calculations.[19,20]

Core-correlation effects were obtained from conventional all-electron CCSD(T)/cc-pCVTZ energies relative to frozen-core CCSD(T)/cc-pCVTZ results and were added to the CCSD(T)-F12a/vtz-f12 energies. Likewise, high-order correlation effects were determined from CCSDT(Q)/cc-pVTZ calculations in comparison to CCSD(T)/cc-pVTZ results.[21] For performing the CCSDT(Q)/cc-pVTZ calculations, we have used Kallays's string-based many-body Mrcc program[22,23] interfaced to the Molpro suite of *ab initio* programs.[10] In order to account for diagonal Born−Oppenheimer effects, these were computed at the CCSD/cc-pVTZ level, as described in detail by Gauss et al.[24] These calculations were performed with the Cfour program package.[25] Scalar relativistic effects were investigated using the Douglas−Kroll−Hess one-electron Hamiltonian,[26] while spin−orbit effects were not explicitly accounted for as they were estimated as being too small. Note that borane is a closed-shell molecule giving rise to second order spin−orbit effects only and that the spin−orbit splitting of the borane atom, i.e., at the dissociation limit of the borane molecule, is 16 $cm^{-1}$ only. As we consider only a small fraction of the global potential energy surface near the equilibrium structure, we believe that the corrections to the vibrational frequencies are below 0.1 $cm^{-1}$ and thus significantly lower than for example errors in the fitting procedure.

**B. Vibrational Structure Calculations.** The expansion of the potential in terms of normal coordinates was truncated after the three-mode contributions. Core-correlation effects, high-order corrections to the coupled-cluster expansion, and scalar relativistic and adiabatic effects were added to the one-mode terms only. This essentially corresponds to a multilevel approximation,[27,28] which has successfully been used by several groups.[29,30] The potential has been generated in an automated fashion using an iterative interpolation algorithm.[13] Initially, 20 grid points have been generated along each normal coordinate. From these, single particle potentials were determined in a subsequent step. In the Watson Hamiltonian, vibrational angular momentum terms were accounted for up to the first order in a multimode expansion of the $\mu$ tensor.[31] Modals were obtained from a VSCF algorithm on the basis of a discrete variable representation (DVR) and a distributed Gaussian basis.[32,33] Vibration correlation effects were accounted for by state-specific VCI calculations including single to quadruple excitations.[14] No more than 10 431 configurations were considered in the correlation space.

## III. Impact of Corrections

The success of frozen-core CCSD(T) calculations is at least partially due to an error compensation of core−valence effects and high-order correlation terms. It is common knowledge that explicitly accounting for just one of these effects usually leads to an unbalanced treatment and thus to worse results. However, this knowledge is primarily based on single point calculations at stationary points or—in the context of calculating accurate vibrational transitions—on two-atomic molecules.[34] With respect to extended parts of potential energy surfaces of polyatomic molecules, knowledge is rather limited.[35] For that reason, we have plotted the core-correlation and high-order corrections along the one-dimensional elongations for all modes in Figure 1. According to Figure 1, core-correlation effects are more important than high-order corrections for all modes. However, for electronically more demanding systems with more valence electrons and multiple bonds between non-hydrogen atoms, this may change. Although core-correlation and high-order corrections partially cancel out for some of the modes, this does not hold true for others. For example, both core-correlation and high-order effects correct the potential related to mode $\nu_3$ in the same direction. As a consequence, error compensation effects are smaller for this molecule than originally anticipated. It is obvious from the shape of correction potentials that core-correlation and high-order terms are more important for an accurate description of high-lying vibrational states than the low-lying fundamental modes. Table 1 shows CCSD(T)-F12a/vtz-f12 anharmonic frequencies and the corresponding corrections due to core-correlation (CC), high-level terms in the coupled-cluster expansion and scalar relativistic (Rel) and adiabatic (DBOC) effects. The latter two corrections provide almost constant shifts of 1390 and 470 $cm^{-1}$ to the potential energy surface but do hardly affect the transitions considered here; i.e., the variation in the region around the equilibrium structure is very small.

For the deuterated species, these effects are even smaller and, in addition, cancel each other out in most cases. As we consider the error in the fitting of our potential energy surfaces to be larger than these effects, we do not discuss them any further. In contrast to that, core-correlation and high-order effects are significantly larger and thus need to be accounted for in order to obtain accurate results.

Martin and Lee computed the harmonic frequencies at the CCSD(T)/cc-pVQZ level for $^{11}BH_3$ and $^{10}BH_3$ (see Table 2), which are in very nice agreement with our results at the

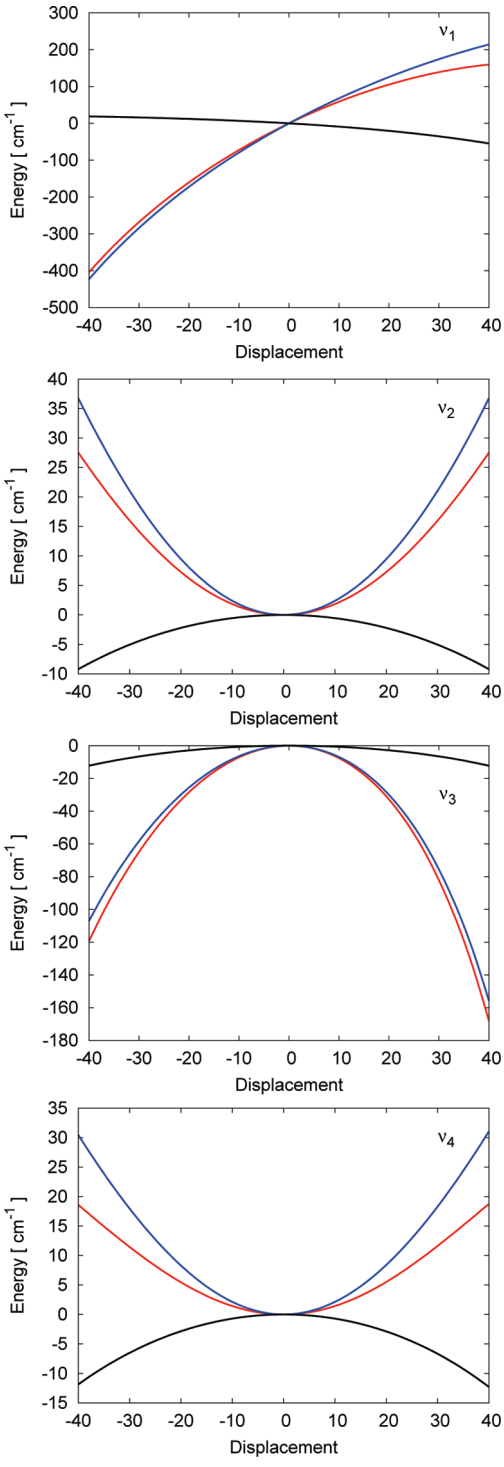



***Figure 1.*** Impact of core-correlation (black) and high-order corrections (black) on the 1D potentials of $BH_3$. The resulting correction is shown in red; the order of the potentials is $\nu_1$ (uppermost), $\nu_2$, $\nu_3$, and $\nu_4$ (bottom).

CCSD(T)-F12a/vtz-f12 level, i.e., $\omega_1 = 2568.0$ cm$^{-1}$, $\omega_2 = 1156.6$ cm$^{-1}$, $\omega_3 = 2701.1$ cm$^{-1}$, and $\omega_4 = 1218.3$ cm$^{-1}$ for $^{11}BH_3$.

The largest deviation is thus 1.7 cm$^{-1}$, which may either arise from the incompleteness of the cc-pVQZ basis used by Martin and Lee or the numerical noise caused by the auxiliary basis sets in the explicitly correlated coupled-cluster calculations. However, corrections due to core-correlation and high-order terms are significantly larger (up to 5.7 cm$^{-1}$) and thus are mainly responsible for the differences between the data sets of Martin and Lee and ours. Note that two modes ($\omega_1$, $\omega_3$) are red-shifted by the corrections, while the other two ($\omega_2$, $\omega_4$) are blue-shifted; i.e., there appears not to be a general tendency.

***Table 1.*** Impact of Core-Correlation Effects, High-Order Coupled-Cluster Terms, Relativistic Contributions, and Diagonal Born−Oppenheimer Corrections on the Fundamental Modes of $^{11}BH_3$ (All Quantities Are Given in cm$^{-1}$)

| | $\nu_i^{\mathrm{F12},a}$ | $\Delta\nu_i^{\mathrm{CC},b}$ | $\Delta\nu_i^{\mathrm{HLT},c}$ | $\Delta\nu_i^{\mathrm{Rel},d}$ | $\Delta\nu_i^{\mathrm{DBOC},e}$ |
|---|---|---|---|---|---|
| $\nu_1$ | 2494.4 | 5.9 | −2.4 | 0.1 | −0.7 |
| $\nu_2$ | 1146.2 | 1.5 | −0.7 | 0.3 | −0.2 |
| $\nu_3$ | 2597.0 | 5.9 | −2.1 | −0.1 | −0.6 |
| $\nu_4$ | 1194.9 | 1.9 | −0.7 | 0.2 | −0.1 |

[a] CCSD(T)-F12a/vtz-f12 anharmonic frequencies. [b] Correction due to core correlation. [c] Correction due to high-level terms in the coupled-cluster expansion. [d] Scalar relativistic corrections. [e] Diagonal Born−Oppenheimer correction.

***Table 2.*** Harmonic and Anharmonic Vibrational Frequencies of $^{11}BH_3$ in cm$^{-1}$

| | ML[a]/Schwenke[b] | | | this work | | |
|---|---|---|---|---|---|---|
| | $\omega_i^a$ | $\nu_i^a$ | $\nu_i^b$ | $\omega_i$ | $\nu_i$ | exp. |
| $\nu_1$ | 2568.3 | 2494.9 | 2491.9 | 2562.4 | 2497.4 | |
| $\nu_2$ | 1158.0 | 1134.2 | 1148.1 | 1160.0 | 1147.1 | 1147.50 |
| $\nu_3$ | 2700.3 | 2587.5 | 2591.5 | 2695.5 | 2600.3 | 2601.57 |
| $\nu_4$ | 1220.0 | 1196.4 | 1199.1 | 1220.9 | 1196.3 | 1196.66 |
| $2\nu_2$ | | | 2277.8 | | 2277.1 | |
| $2\nu_4^0$ | | 2361.5 | 2381.2 | | 2383.6 | |
| $\nu_1 + \nu_2$ | | | | | 3650.3 | |
| $\nu_1 + \nu_4$ | | | | | 3685.2 | |
| $\nu_2 + \nu_3$ | | | | | 3736.3 | |
| $\nu_2 + \nu_4$ | | | 2354.9 | | 2351.4 | |

[a] Data taken from ref 5. [b] Data taken from ref 9.

## IV. Results and Discussion

The BH bond length was determined to be $r_e = 1.1895$ Å at the CCSD(T)-F12a/vtz-f12 level, which is in excellent agreement with the CCSD(T)/cc-pVQZ value of $r_e = 1.1899$ Å obtained by Martin and Lee.[5] Both values are slightly larger than that determined from the experimental data of Kawaguchi,[3] i.e., $r_e = 1.185$ Å. However, the latter value was estimated from an $r_0$ bond length, which has subsequently been corrected to $r_e$ by an increment taken from the $CH_3^+$ cation.

Our computed harmonic and anharmonic frequencies are summarized for all isotopologues in Tables 2−4. The agreement of our computed fundamentals with the experimental data of Kawaguchi et al.[2−4] is excellent, with a mean absolute deviation of just 0.9 cm$^{-1}$. We believe that the remaining errors arise mainly from the fitting procedure and the truncation of the potential after the three-mode terms in the vibrational structure calculations.

Although a direct comparison of our computed values for vibrational overtones and combination bands with experimental data is not possible, we expect slightly larger errors for these modes. A comparison with the data of Kaldor and Porter[1] or Tague and Andrews[36] has not been provided, as these data are affected by matrix isolation effects, which can be quite substantial for these light molecules.

***Table 3.*** Harmonic and Anharmonic Vibrational Frequencies of $^{10}BH_3$ in $cm^{-1}$

| | ML[a] | | this work | | |
|---|---|---|---|---|---|
| | $\omega_i$ | $\nu_i$ | $\omega_i$ | $\nu_i$ | exp. |
| $\nu_1$ | 2568.3 | 2498.4 | 2562.4 | 2500.3 | |
| $\nu_2$ | 1170.3 | 1145.9 | 1172.4 | 1159.0 | |
| $\nu_3$ | 2716.0 | 2601.6 | 2711.2 | 2614.5 | 2615.79 |
| $\nu_4$ | 1225.9 | 1202.0 | 1226.8 | 1201.3 | |
| $2\nu_2$ | | | | 2297.6 | |
| $2\nu_4^0$ | | 2367.6 | | 2394.7 | |
| $\nu_1 + \nu_2$ | | | | 3667.3 | |
| $\nu_1 + \nu_4$ | | | | 3695.5 | |
| $\nu_2 + \nu_3$ | | | | 3762.1 | |
| $\nu_2 + \nu_4$ | | | | 2369.1 | |

[a] Data taken from ref 5.

***Table 4.*** Harmonic and Anharmonic Vibrational Frequencies of $^{10}BD_3$ and $^{11}BD_3$ in $cm^{-1}$

| | $^{10}BD_3$ | | | $^{11}BD_3$ | | |
|---|---|---|---|---|---|---|
| | GVS[a] | this work | | GVS[a] | this work | |
| | $\omega_i$ | $\omega_i$ | $\nu_i$ | $\omega_i$ | $\omega_i$ | $\nu_i$ |
| $\nu_1$ | 1813 | 1812.3 | 1731.6 | 1813 | 1812.3 | 1724.8 |
| $\nu_2$ | 936 | 920.5 | 910.5 | 920 | 904.7 | 895.1 |
| $\nu_3$ | 2040 | 2038.7 | 1984.2 | 2018 | 2016.9 | 1963.9 |
| $\nu_4$ | 916 | 905.7 | 890.8 | 910 | 899.8 | 885.1 |
| $2\nu_2$ | | | 1841.7 | | | 1820.7 |
| $2\nu_4^0$ | | | 1778.2 | | | 1766.7 |
| $\nu_1 + \nu_2$ | | | 2619.9 | | | 2596.9 |
| $\nu_1 + \nu_4$ | | | 2606.6 | | | 2595.0 |
| $\nu_2 + \nu_3$ | | | 2887.3 | | | 2852.0 |
| $\nu_2 + \nu_4$ | | | 1804.7 | | | 1783.8 |

[a] CISDTQ/TZ2P data taken from ref 7.

The largest difference between the data sets of Martin and Lee,[5] Schwenke,[9] and ourselves is observed for $\nu_3$ of $^{10}BH_3$ *and* $^{11}BH_3$. In both cases, the values relying on the quartic force field appear to be too low. Therefore, we assume that this is mainly an effect arising from the approximation of the PES. Moreover, our value for the $2\nu_4^0$ overtone of $^{11}BH_3$ is in nice agreement with the value of Schwenke but differs considerably from Martin and Lee's value. Therefore, we believe that Martin and Lee's value for the corresponding overtone in $^{10}BH_3$ is too low. Anharmonic frequencies for the perdeuterated species are provided for the first time and can thus not be compared. However, we consider the anharmonic frequencies predicted for these system to be accurate enough in order to guide experiments to come. Zero point vibrational energies (ZPVE) have been determined for all four isotopologues: 5715.5 $cm^{-1}$ ($^{11}BH_3$), 5742.6 $cm^{-1}$ ($^{10}BH_3$), 4242.9 $cm^{-1}$ ($^{11}BD_3$), and 4277.9 $cm^{-1}$ ($^{10}BD_3$). Our *ab initio* value for the ZPVE of $^{11}BH_3$ essentially is identical with the value of Feller et al.[8] (5715.6 $cm^{-1}$) obtained from a combined experimental/theoretical scaling procedure.

We like to state once more that the calculations including core-correlation effects and high-order terms have been performed in a fully automated fashion, controlled by some very few keywords in the input section only. Consequently, this study shows that modern *ab initio* programs are able to predict highly accurate vibrational spectra without extensive fitting etc. by hand.

## V. Summary and Conclusions

Vibrational frequencies of borane and its isotopologues have been determined from multidimensional potential energy surfaces obtained from high-level *ab initio* calculations. Mean absolute deviations were found to be below one wavenumber with respect to the most accurate gas phase measurements. Likewise, zero point vibrational energies, which are important for an accurate calculation of heats of formation, were found to be in excellent agreement with previous data. Several vibrational transitions were provided for the first time. Moreover, core-correlation effects and high-order contributions in the coupled-cluster treatment do not cancel out for all types of vibrations and were thus found to be important to be included in the calculations. Scalar relativistic effects and diagonal Born−Oppenheimer corrections were found to be negligible.

**Acknowledgment.** We thank Prof. T. Hrenar (University of Zagreb) for performing the DBOC calculations for us. Financial support by the Deutsche Forschungsgemeinschaft is kindly acknowledged.

### References

(1) Kaldor, A.; Porter, R. *J. Am. Chem. Soc.* **1971**, *93*, 2140.

(2) Kawaguchi, K.; Butler, J. E.; Yamada, C.; Bauer, S. H.; Minowa, T.; Kanamori, H.; Hirota, E. *J. Chem. Phys.* **1987**, *87*, 2438.

(3) Kawaguchi, K. *J. Chem. Phys.* **1992**, *96*, 3411.

(4) Kawaguchi, K. *Can. J. Phys.* **1994**, *72*, 925.

(5) Martin, J. M. L.; Lee, T. J. *Chem. Phys. Lett.* **1992**, *200*, 502.

(6) Botschwina, P. In *Ion and cluster spectroscopy and structure*; Maier, J. P., Ed.; Elsevier: Amsterdam, 1989; p 59.

(7) Galbraith, J. M.; Vacek, G.; Schaefer, H. F., III. *J. Mol. Struct.* **1993**, *300*, 281.

(8) Feller, D.; Dixon, D. A.; Peterson, K. A. *J. Phys. Chem. A* **1998**, *102*, 7053.

(9) Schwenke, D. W. *J. Phys. Chem.* **1996**, *100*, 2867.

(10) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. Molpro, development version 2009 3, a package of ab initio programs (2008). http://www.molpro.net (accessed Nov 2010).

(11) Bowman, J. *Acc. Chem. Res.* **1986**, *19*, 202.

(12) Gerber, R.; Ratner, M. *Adv. Chem. Phys.* **1988**, *70*, 97.

(13) Rauhut, G. *J. Chem. Phys.* **2004**, *121*, 9313.

(14) Neff, M.; Rauhut, G. *J. Chem. Phys.* **2009**, *131*, 124129.

(15) Peterson, K.; Adler, T.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.

(16) Rauhut, G.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054105.

(17) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285.

(18) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175.

(19) Adler, T.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 034106.

(20) Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.

(21) Kallay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 214105.

(22) Kallay M. Ph. D. Thesis, Budapest University of Technology and Economics, Budapest, Hungary, 2001.

(23) Kallay, M.; Surjan, P. *J. Chem. Phys.* **2001**, *115*, 2945.

(24) Gauss, J.; Tajti, A.; Kallay, M.; Stanton, J. F.; Szalay, P. G. *J. Chem. Phys.* **2006**, *125*, 144111.

(25) *CFOUR*, Coupled-Cluster techniques for Computational Chemistry, a quantum-chemical program package by Stanton, J. F.; Gauss, J.; Harding, M. E.; Szalay, P. G. with contributions from Auer, A.A.; Bartlett, R. J.; Benedikt, U.; Berger, C.; Bernholdt, D. E.; Bomble, Y. J.; Cheng, L.; Christiansen, O.; Heckert, M.; Heun, O.; Huber, C.; Jagau, T.-C.; Jonsson, D.; Juselius, J.; Klein, K.; Lauderdale, W. J.; Matthews, D. A.; Metzroth, T.; O'Neill, D. P.; Price, D. R.; Prochnow, E.; Ruud, K.; Schiffmann, F.; Schwalbach, W.; Stopkowicz, S.; Tajti, A.; Vazquez, J.; Wang, F.; Watts, J. D. And the integral packages *MOLECULE* (Almlof, J.; Taylor, P. R.), *PROPS* (Taylor, P. R.), *ABACUS* (Helgaker, T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.), and ECP routines by Mitin, A. V.; van Wullen, C. For the current version, see http://www.cfour.de (accessed Nov 2010).

(26) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89.

(27) Pflüger, K.; Paulus, M.; Jagiella, S.; Burkert, T.; Rauhut, G. *Theor. Chim. Acta* **2005**, *114*, 327.

(28) Rauhut, G.; Barone, V.; Schwerdtfeger, P. *J. Chem. Phys.* **2006**, *125*, 054308.

(29) Yagi, K.; Hirata, S.; Hirao, K. *Theor. Chim. Acta* **2007**, *118*, 681.

(30) Sparta, M.; Høyvik, I.-M.; Toffoli, D.; Christiansen, O. *J. Phys. Chem. A* **2009**, *113*, 8712.

(31) Watson, J. K. G. *Mol. Phys.* **1968**, *15*, 479.

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

(33) Hamilton, I.; Light, J. *J. Chem. Phys.* **1986**, *84*, 306.

(34) Ruden, T. A.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2004**, *121*, 5874.

(35) Koput, J.; Peterson, K. A. *J. Chem. Phys.* **2006**, *125*, 044306.

(36) Tague, T.; Andrews, J. *J. Am. Chem. Soc.* **1994**, *116*, 4970.

CT1004752

JCTC Journal of Chemical Theory and Computation

# Multiconfigurational Second-Order Perturbation Theory Restricted Active Space (RASPT2) Method for Electronic Excited States: A Benchmark Study

Vicenta Sauri,[†] Luis Serrano-Andrés,[†,⊥] Abdul Rehaman Moughal Shahi,[‡] Laura Gagliardi,*[,§] Steven Vancoillie,[||] and Kristine Pierloot*[,||]

*Instituto de Ciencia Molecular, Universitat de València, P.O. Box 22085, ES-46071 Valencia, Spain, Department of Physical Chemistry, University of Geneva, 30, q. E. Ansermet, 1211 Genève, Switzerland, Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota 55455-0431, United States, and Department of Chemistry, Katholieke Universiteit Leuven, Belgium*

Received August 24, 2010

**Abstract:** The recently developed second-order perturbation theory restricted active space (RASPT2) method has been benchmarked versus the well-established complete active space (CASPT2) approach. Vertical excitation energies for valence and Rydberg excited states of different groups of organic (polyenes, acenes, heterocycles, azabenzenes, nucleobases, and free base porphin) and inorganic (nickel atom and copper tetrachloride dianion) molecules have been computed at the RASPT2 and multistate (MS) RASPT2 levels using different reference spaces and compared with CASPT2, CCSD, and experimental data in order to set the accuracy of the approach, which extends the applicability of multiconfigurational perturbation theory to much larger and complex systems than previously. Relevant aspects in multiconfigurational excited state quantum chemistry such as the valence−Rydberg mixing problem in organic molecules or the double d-shell effect for first-row transition metals have also been addressed.

## Introduction

The development and efficient implementation of the multiconfigurational second-order perturbation theory approach (CASPT2) in the beginning of the 1990s by Roos and coworkers[1] represented a breakthrough for quantum chemistry in general, but more specifically for the study of those electronic structure cases which required a multiconfigurational description of the reference wave function. Those problems were then first solved quantitatively or accurately in many polyatomic systems, such as a number of bond breakings and dissociations,[2] potential energy hypersurface (PEH) degeneracies (conical intersections),[3] symmetry breaking problems (Cope rearrangement),[4] biradical situations,[5,6] organic molecules photophysics,[7−10] transition metal (TM) bonding[11−16] and spectroscopy,[17−22] and actinide chemistry.[23−26]

In particular, this method showed to be best suited to deal with the quantum chemistry of the excited state.[27] For the first time, the overall level of accuracy in the determination of electronic excitation energies in small- to medium-sized molecules reached 0.1−0.3 eV for systems up to 30 atoms, like free base porphin.[28] The ability of a multiconfigurational approach that extensively includes correlation effects like the CASPT2 method opened the door for studying spectroscopic and photochemical phenomena in systems in which computationally more costly approaches such as multireference configuration interaction (MRCI) could not be applied. In a recent benchmark study of excitation energies in organic molecules,[29] the current version of the CASPT2 method—in which the IPEA zeroth-order Hamiltonian[30] was used—was

* Corresponding authors. E-mail: gagliard@umn.edu (L.G.); Kristin.Pierloot@chem.kuleuven.be (K.P.).
[†] Universitat de València.
[‡] University of Geneva.
[§] University of Minnesota.
[||] Katholieke Universiteit Leuven.
[⊥] Deceased, September 2010.

10.1021/ct100478d 
Published on Web 12/08/2010

shown to be superior to the CCSD and CC2 approaches and equally as accurate as the linear response coupled-cluster CC3 procedure for closed-shell ground-state systems. CASPT2 has also proven to be much better than any other method in medium-sized molecules for excited states in cases where the ground state is poorly defined by a closed-shell configuration (biradicals, conical intersections, dissociation paths) or the excited state has a strong multiconfigurational or diexcited character.[29,31]

The CASPT2 method is, however, not free of drawbacks, all of them actually related with the limitation in the size of the complete active space (CAS) due to the difficulties in handling large high-order density matrices. Some of the weaknesses of the CASPT2 method which were partially solved were (i) the intruder state problem perturbing the first-order interacting reference space—which required the introduction of a level-shift correction (LS-CASPT2);[32,33] (ii) the overestimation of high-multiplicity states—solved by the IPEA zeroth-order Hamiltonian;[30] (iii) the lack of orthogonality of the individual single-state CASPT2 solutions, causing improper mixing of valence and Rydberg wave functions[7] or unphysical state crossings[34]—which was solved by developing the multistate (MS) CASPT2 approach.[35] These weaknesses generally affected the size and accuracy of the problems under study because the maximum active space size of 13−15 molecular orbitals (MOs) was rapidly reached, especially if a large number of Rydberg orbitals has to be considered. Specific problems like, for instance, the need to include in the CAS a second d-shell (4d) for first-row transition metal atoms were also soon recognized.[36]

As mentioned, all of these problems have one straightforward solution: increasing the active space size. Recent methodological developments in *ab initio* quantum chemistry like the Cholesky decomposition (CD) approach, also implemented for the CASPT2 method,[37,38] have reduced the effort involved in handling two-electron integrals and have extended the applicability of *ab initio* methods to much larger molecular systems. For instance, a 1000 basis set CD-CASPT2 calculation is easily affordable nowadays within the MOLCAS package.[38−41] Truncation of the virtual space in CD-CASPT2 calculations is also possible.[42] This method is based on a modified version of the frozen natural orbital (FNO) approach used in coupled cluster theory. However, increasing the molecular size or truncating the virtual space are not always sufficient techniques. There are many quantum-chemical problems whose accurate solution is out of reach for CASPT2 because the adequate number of MOs cannot be included in the active space, like organometallic systems with more than one transition metal atom, extended $\pi$-space molecules with more than 14 $\pi$ MOs, problems in which the inclusion of additional $\sigma$ MOs is needed, or chemical reactions in which the requirements for a balanced space exceeds the capability of the CAS approach.

The recent implementation of the second-order perturbation theory restricted active space (RASPT2) method leads the field in the proper direction.[43] In the CAS framework,[44] the MO space is divided into three subspaces with a varying number of electrons: inactive (always doubly occupied), active (with varying occupation from zero to two), and secondary (always empty). All possible excitation levels compatible with spatial and spin symmetry involving the electrons in the active space form the multiconfigurational CAS-CI space used as a reference for a further perturbative CASPT2 treatment.[1,2] The configurational space rapidly grows to many millions of determinants with the size of the active space, making the treatment unaffordable. The RAS method[45,46] further divides the active space into three subspaces: RAS1, RAS2, and RAS3. The final multireference space is built by allowing in RAS2 the same type of full-CI expansion as previously in CAS, but restricting in RAS1 and RAS3 the excitation level to a predefined range: up to single (S), double (SD), triple (SDT), quadruple (SDTQ), etc., by limiting the number of allowed holes (RAS1) and particles (RAS3). Avoiding high excitation levels in RAS1 and RAS3 leads to less extended multiconfigurational spaces, therefore allowing a much larger number of active orbitals as compared with the CAS expansions. However, the number of possible divisions of the active space combined with the different allowed levels of excitation increases the choices of configurational expansions and the number of solutions, thus making the RASSCF/RASPT2 method less systematic than the CASSCF/CASPT2 method.

The selection of RAS spaces requires very careful calibration. Finding reliable strategies for general purpose calculations is highly required. Up to now, the RASPT2 method has been tested in the determination of the singlet−triplet state energy splitting of three copper−dioxygen and two copper−oxo complexes[43,47] and one-electron ionization potential and optical band gaps of ethylene, acetylene, and phenylene oligomers.[48] It has been shown how RASPT2 offers a similar accuracy when compared to CASPT2 at significantly reduced computational expense, whereas more demanding calculations out of reach for CASPT2 can be performed with the new formulation.

In the present contribution, we focus on electronic excitation energies, a field in which RASPT2 probably will play a major role in the coming years. We have selected different sets of molecules in order to check the accuracy of the method and the computational strategies in systems and problems of various classes, including valence and Rydberg, singlet and triplet, and ligand-field and charge-transfer excited states in different organic and inorganic systems. Figure 1 compiles the benchmark set of molecules considered in the present study.

The paper is divided into four sections, each of which focuses on one aspect of the calibration. Initially, free base porphin will be used as an example of the use of RASPT2 in a system with an extended $\pi$ system whose inclusion in the active space is out of reach for CASPT2. Then, excitation energies of singlet and triplet valence and Rydberg states of ethene and benzene will be computed in order to test the accuracy and ability of RASPT2 to deal with the simultaneous calculation of valence and Rydberg states, and how it takes care of the valence−Rydberg problem. Next, the valence states of naphthalene and a set of organic five- and six-membered heterocyclic molecules and DNA/RNA nucleobases will be computed with RASPT2 to determine the accuracy of the results with different partitions of the RAS

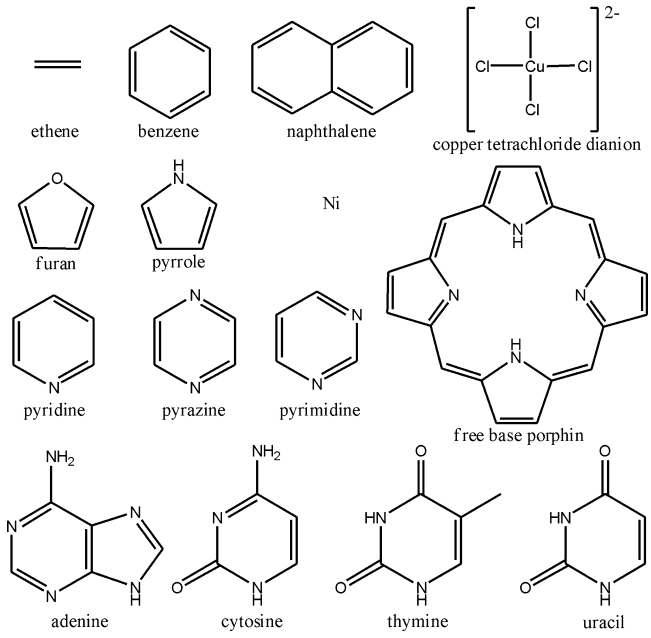



***Figure 1.*** Benchmark set of molecules considered in this study.

space. Finally, the nickel atom and the copper tetrachloride dianion will be computed in order to establish the accuracy and proper strategies required in RASPT2 to handle the required inclusion of a second correlating d shell for first-row transition metal compounds, and how the new method simplifies the calculations and extends their possibilities.

## 2. Computational Details

All CASSCF/CASPT2 and RASSCF/RASPT2 calculations were performed with the MOLCAS-7 program.[39] All-electron, one-electron basis sets were used throughout. For ethene and benzene (and except when indicated), the calculations were performed with an ANO-L basis set[49] contracted to [4s3p1d] for carbon and [2s1p] for hydrogen atoms. In addition, 1s-type, 1p-type, and 1d-type contracted functions, with diffuse coefficients described elsewhere,[35] were added to this basis set and placed in the center of the molecule to describe Rydberg orbitals. For free base porphin, an ANO-S C,N[3s2p1d]/H[2s] basis set was used. For all other organic systems, a triple-$\zeta$ valence polarized basis (TZVP) set was employed.[50] Cholesky decomposition of the two-electron integrals was accomplished with a threshold of $10^{-5}$ au.[51] Reduced-scaling evaluation of the Fock exchange matrices in the CASSCF and RASSCF calculations was accomplished by means of the Local-K screening approach[52] employing localized Cholesky orbitals.[53] For the calculations on transition metal systems, ANO-RCC basis sets[54] were used, contracted to [7s6p4d3f2g] for nickel, [7s6p4d3f2g1h] for copper, and [5s4p2d1f] for chlorine atoms. Scalar relativistic effects were included using a Douglas−Kroll−Hess Hamiltonian.[55,56] In ethene and benzene, except when mentioned, the ground state geometries were taken from gas-phase experimental determinations, as described in the Supporting Information (SI). $CuCl_4^{2-}$ is square-planar ($D_{4h}$), and the Cu−Cl distance, 2.291 Å, was taken from our previous study.[57] The other systems were optimized with density functional theory (DFT) and the B3LYP functional[58] using the Turbomole 5.10 package and employing the triple-$\zeta$ valence polarized basis sets available in Turbomole.[59] At the optimized geometries, subsequent single point MS-CASSCF/CASPT2 and MS-RASSCF/RASPT2 calculations were performed using the mentioned basis sets. When high symmetry is required, for instance, $D_{6h}$ in benzene, $D_{4h}$ for $CuCl_4^{2-}$, or spherical symmetry in the nickel atom, the calculations were performed in a lower-symmetry point group, and MOLCAS tools were used to obtain the proper orbital symmetry. An imaginary level shift[32] of 0.1 au was used to prevent weakly coupling intruder states' interference, and the default shift for the IPEA zeroth-order Hamiltonian[30] (0.25 au) was employed. The value of 0.25 au is now the default in CASPT2, and we have used it in all cases, except for the free base porphin calculations where we have set the shift equal to zero. The reason for this different choice is that we wanted to compare the present results to those obtained prior to the introduction of the IPEA shift in 2004.[28,60] However this study does not focus on IPEA shift effects, and we do not compare results with different IPEA shifts for each system because this deviates from the purpose of the present study. In all calculations, the core electrons are kept frozen in the perturbative calculations, except for the nickel atom where the 3s and 3p electrons are included.

It is important to describe the notation employed to label CAS and RAS calculations. In the first case, the traditional label is used, that is, CAS($n$,$i$), where $n$ is the number of electrons included in the active space and $i$ is the number of active orbitals. For RAS calculations, a longer notation is used, RAS($n$,$l$,$m$;$i$,$j$,$k$), where $n$ is the number of active electrons, $l$ the maximum number of holes allowed in RAS1, and $m$ the maximum number of electrons to enter in RAS3. Active orbitals are labeled by $i$,$j$,$k$ and refer to those placed in RAS1, RAS2, and RAS3, respectively. Sometimes, we will also use S, SD, SDT, or SDTQ to emphasize the maximum RAS1→RAS3 excitation level. The RAS2 subspace has the same meaning in RASSCF as the CAS active space in a CASSCF calculation; i.e., all possible spin- and spatial symmetry-adapted configuration state functions (CSFs) that can be constructed from the orbitals in RAS2 are included in the multiconfigurational wave function. The RAS1 and RAS3 subspaces, on the other hand, permit the generation of additional CSFs subject to the restriction that a limited number of excitations may occur from RAS1, which otherwise contains only doubly occupied orbitals, and a limited number of excitations may occur in RAS3, which otherwise contains only empty orbitals. The active space employed for the various systems will be described in each section. Further details are reported in the SI .

## 3. Results and discussion

**3.A. Free Base Porphin.** Free base porphin (FBP) is an example of an extended $\pi$-conjugated system having 26 valence $\pi\pi^*$ electrons and 24 $\pi\pi^*$ MOs (26/24). Including a full $\pi\pi^*$ active space is out of reach for a conventional CASPT2 calculation. Previous studies were performed at the CASPT2(4/4) and CASPT2(16/14) levels.[28,60] In the former

***Table 1.*** Excitation Energies (eV) of the Singlet and Triplet Valence $\pi\pi^*$ States of Free Base Porphin ($D_{2h}$)

| state | CASPT2 (4/4)[a] | CASPT2 (16/14)[b] | RASPT2 (26,2,2;11,4,9) (SD)[c] | RASPT2 (26,3,3;11,4,9) (SDT)[c] | RASPT2 (26,2,2;i,j,k) (SD)[d] | STEOM-CCSD[e] | exptl[f] |
|---|---|---|---|---|---|---|---|
| $1^1B_{3u}$ | 1.70 | 1.63 | 2.18 | 1.91 | 2.18 | 1.75 | 1.98 − 2.02 ($Q_x$) |
| $1^1B_{2u}$ | 2.26 | 2.11 | 2.38 | 2.16 | 2.38 | 2.40 | 2.33 − 2.42 ($Q_y$) |
| $2^1B_{2u}$ | 2.91 | 3.08 | 3.23 | 2.86 | 3.23 | 3.62 | 3.13 − 3.33 (B) |
| $2^1B_{3u}$ | 3.04 | 3.12 | 3.21 | 3.16 | 3.21 | 3.47 | 3.13 − 3.33 (B) |
| $3^1B_{2u}$ | | 3.42 | 5.22 (3.30)[g] | 3.37 | 3.80 | 4.35 | 3.65 (N) |
| $3^1B_{3u}$ | | 3.53 | 5.38 (3.21)[g] | 3.28 | 3.48 | 4.06 | 3.65 (N) |
| $4^1B_{2u}$ | | 3.96 | 5.95 (4.02)[g] | 4.10 | 4.20 | 5.00 | 4.25 (L) |
| $4^1B_{3u}$ | | 4.04 | 6.04 (4.14)[g] | 4.22 | 4.23 | 5.17 | 4.25 (L) |
| $1^3B_{2u}$ | | 1.52 | 1.83 | 1.70 | 1.83 | 1.26 | 1.58 |
| $1^3B_{3u}$ | | 1.85 | 1.99 | 1.77 | 1.99 | 1.80 | |
| $2^3B_{3u}$ | | 1.88 | 1.98 | 1.88 | 1.98 | 1.98 | |
| $2^3B_{2u}$ | | 1.98 | 1.98 | 1.90 | 1.98 | 1.85 | |
| CSF[h] | 8 | 537705 | 63258 (1877432)[g] | 1877565 | 279974/676297 | | |

[a] CASPT2(4,4)/ANO-L 3s2p/2s, ref 60. Gouterman's four-electron/four-MO CAS space. [b] CASPT2(16,14)/ANO-S 3s2p1d/2s, ref 28. [c] Present RASPT2 results. Full $\pi\pi^*$ 26-electron/24-MO RAS employed. Gouterman's 4/4 space placed in RAS2. SD or SDT for all states except when indicated. The poor results for the highest singlet states explained in the text. [d] Different RAS spaces partition following the occupation number criterion of Table 2. See text. [e] STEOM-CCSD/SVZP results from ref 61. [f] See data in ref 28. [g] Present RASPT2 results. Full $\pi\pi^*$ 26-electron/24 MO-RAS employed. Gouterman's 4/4 space placed in RAS2. Within parentheses are results using SD for the ground $1^1A_g$ state and SDT for the excited state and CSFs for the excited-state $^1B_{2u}$ SDT calculations. [h] Number of configuration state functions (CSF) for the $^1A_g$ symmetry. In the sixth column are CSFs for active spaces (26,2,2;8,6,9)(SD)/(26,2,2;6,8,9)(SD) as examples.

***Table 2.*** Natural Occupation Numbers of the Most Relevant Molecular Orbitals of the Low-Lying $\pi\pi^*$ States of Free Base Porphin ($D_{2h}$)[a]

| state | $3b_{1u}$ | $4b_{1u}$ | $2b_{2g}$ | $3b_{2g}$ | $3b_{3g}$ | $5b_{1u}$[b] | $2a_u$[b] | $4b_{2g}$[b] | $4b_{3g}$[b] | $3a_u$ | RAS2[c] $b_{1u}b_{2\ g}b_{3\ g}a_u/e^-$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $1^1A_g$ | 1.9692 | 1.9596 | 1.9714 | 1.9561 | 1.9581 | 1.8539 | 1.8631 | 0.1481 | 0.1595 | 0.0543 | |
| $1^1B_{3u}$ | 1.9699 | 1.9606 | 1.9349 | 1.9715 | 1.9646 | 1.4726 | 1.4449 | 0.5448 | 0.5664 | 0.0739 | 1111/4 |
| $1^1B_{2u}$ | 1.9696 | 1.9622 | 1.9702 | 1.9601 | 1.9625 | 1.5462 | 1.3962 | 0.6117 | 0.4803 | 0.0626 | 1111/4 |
| $2^1B_{2u}$ | 1.9609 | 1.9684 | 1.9662 | 1.9582 | 1.9638 | 1.3169 | 1.4744 | 0.5389 | 0.6988 | 0.0615 | 1111/4 |
| $2^1B_{3u}$ | 1.9713 | 1.9527 | 1.9526 | 1.9722 | 1.9451 | 1.3941 | 1.4197 | 0.6275 | 0.6037 | 0.0602 | 1111/4 |
| $3^1B_{2u}$ | 1.9906 | 1.2519 | 1.9907 | 1.9637 | 1.7919 | 1.7860 | 1.7749 | 0.2467 | 1.0048 | 0.0992 | 2121/8 |
| $3^1B_{3u}$ | 1.9922 | 1.3057 | 1.9617 | 1.9923 | 1.7339 | 1.8496 | 1.6622 | 1.1639 | 0.1479 | 0.0773 | 2121/8 |
| $4^1B_{2u}$ | 1.2515 | 1.9914 | 1.7929 | 1.9747 | 1.9898 | 1.8367 | 1.7494 | 0.1429 | 1.0928 | 0.0765 | 3311/10 |
| $4^1B_{3u}$ | 1.3231 | 1.9910 | 1.9694 | 1.7753 | 1.9894 | 1.8312 | 1.6278 | 0.9479 | 0.2936 | 0.1310 | 3312/10 |

[a] RASSCF(26,2,2;11,4,9)(SD) level of calculation. [b] Orbitals of the 4/4 Gouterman's space. [c] Orbitals and electrons within RAS2 selected from the occupation numbers (<1.9 and >0.1).

case, only four singlet states were computed, whereas eight singlet and eight triplet states were obtained at the latter level of theory. An overall agreement of 0.2−0.3 eV with respect to experimental values was obtained for all eight singlet states belonging to the porphin Q, B, N, and L bands, although in all these cases, CASPT2 yields too low values. FBP will be employed as a typical example of how to properly select the RAS active spaces and establish a RAS1/RAS3 excitation level yielding balanced and accurate excitation energies.

Table 1 displays a comparison between our new RASPT2 calculations and the previous CASPT2 calculations. The former includes all $\pi\pi^*$ valence electrons and MOs (26/24) in the RAS active space. As in many other organic molecules, the two highest-lying occupied MOs (HOMO and HOMO−1) and the two lowest-lying MOs (LUMO and LUMO+1) are the four most relevant MOs to describe the four lowest-lying states of the molecule.[27] This active space (named Gouterman's space in FBP) was previously used for CASPT2 calculations and showed to be necessary to describe the nature of such states. Indeed, CASPT2(4/4) calculations (see Table 1) provided reasonably accurate values for the mentioned states, as well as CASPT2(16/14), including Gouterman's MOs plus other additional orbitals which allowed calculation of higher roots.

How should we find out how to partition the RAS spaces in order to include FBP full $\pi\pi^*$ space and obtain accurate results? Not all partitions are equally adequate, and especially the choice of RAS2 has to be made carefully. Table 2 summarizes the natural orbital occupation numbers for a number of relevant MOs obtained in a RASSCF(26,2,2;11,4,9)(SD) calculation. This level of theory, including the four Gouterman's MOs in RAS2, the remaining $\pi\pi^*$ occupied and unoccupied MOs in RAS1 and RAS3, respectively, and up to double excitations (SD) for the latter spaces, is not intended to get accurate results for all nine computed states but just to guide us in designing the RAS partition. For each excited state, we have selected (see last column in Table 2) the most relevant MOs, namely, those in which the occupation number is below 1.9 or above 0.1. Obviously such a number may vary at the different RASSCF levels, but just slightly. It is shown, for instance, that for the four lowest-lying states just the four Gouterman's MOs fulfill such requirements, as expected, whereas two more occupied MOs are required for the $3^1B_{2u}$ and $3^1B_{3u}$ states and two and three more for the $4^1B_{2u}$ and $4^1B_{3u}$ states, respectively.

Why is this analysis so important? In order to have a balanced and accurate energy difference between states, the MOs strongly differing in occupation number for such states must be placed in RAS2 simultaneously. That is, to get

balanced RASPT2 excitation energies from the ground to the $1^1B_{2u}$, $2^1B_{2u}$, $1^1B_{3u}$, and $2^1B_{3u}$ excited states, at least the four Gouterman's MOs ($b_{1u}b_{2g}b_{3g}a_u/n_e$;1111/4) must be placed in RAS2. Otherwise the description of the various states will be strongly unbalanced at the initial RASSCF level, and RASPT2 may not be able to recover the desired accuracy. This is better seen in the case of the higher-lying states. In Table 1, we report excitation energies at the RASPT2(26,2,2;11,4,9)(SD) level of calculation. Only the four Gouterman's MOs are included in RAS2, and single and double excitations (SD) are allowed from RAS1 to RAS3 to obtain the RAS-CI expansion. This level is clearly adequate for describing the four lowest-lying states, largely reducing the computational cost with respect to CASPT2(16/14) (only 10% of CSFs required for the RAS calculations). The case is quite different for the four next states, which require additional MOs to be properly described (see Table 2).

The RASPT2 excitation energies deviate toward high values by more than 1.5 eV, showing the underestimation of the correlation energy for the excited states as compared with the ground state. In parentheses, we included the results of increasing the CI excitation level only for the excited states to SDT, while keeping SD for the ground state, a strategy that partially restores the lost balance, giving excitation energies within 0.2−0.3 eV from the experimental values. Similar results are obtained if we increase the level of excitation in RAS1/RAS3 to triple excitations for all states with the RASPT2 (26,3,3;11,4,9)(SDT) calculations, proving that the ground state treatment does not improve with respect to the SD level. In any case, the computational cost increases enormously by including the triple excitations (30 times more CSFs are required).

We also performed more elaborate calculations in which each pair of states (here, the ground and each excited state) has been computed using the specific active space suggested by the occupation numbers in Table 2. In this procedure, both the ground and excited states have in RAS2 those MOs largely changing their occupation number in the excitation process. These calculations (which are equivalent to the RASPT2(26,2,2;11,4,9)(SD) results for the four lowest-lying states) provide the most accurate set of results for the different states at an intermediate computational cost.

The main conclusion obtained from these sets of calculations on FPB is that RASPT2 can provide accurate results for excited states only if the design of the RAS partition, and particularly the composition of the RAS2 space, is carefully controlled. RAS2 must contain those MOs that largely change their occupation number in the states under comparison. Otherwise, the corresponding states will have an unbalanced CI description, and perturbation theory might be unable to provide accurate excitation energies. Any initial RASSCF SD calculation on the requested states including a large enough active space will be sufficient to identify the MOs that should be placed in RAS2, and the occupation number criterion (<1.9 and >0.1) can be used for guidance. If the RAS2 partition is correct, the singles and doubles (SD) level of CI excitation required in RAS1 and RAS3 is sufficient to provide accurate excitation energies at a reasonable computational cost. Increasing the excitation level (triple or quadruple CIs) may partially compensate for the lack of balance, but it typically gives large CI expansions that may become very expensive. The use of the full $\pi\pi^*$ 26/24 active space increases the accuracy compared to more limited active spaces. Furthermore, RASPT2 compares well with experimental results, unlike CCSD, especially for the higher states, which deviate from experimental results by almost 0.8 eV at the CCSD level of calculation. RASSCF/RASPT2 can therefore be considered a very convenient tool for studying the spectrum of this type of $\pi$-extended system. This is even more important when carrying out geometry optimizations. Occasionally, the selective partition of a large $\pi$ space like that of porphin leads to localized solutions at the CASSCF level, which can be avoided at the RASSCF level, as shown recently in psoralen.[62]

**3.B. Ethene and the Valence−Rydberg Mixing Problem.** Ethene is usually described by two $\pi\pi^*$ valence orbitals—the HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital)—that form the basis for the low-lying valence singlet and triplet $\pi\pi^*$ states. Additionally, series of diffuse states of increasing energies converging to the ionization potentials (IPs) of the molecule, named the Rydberg states, will also appear at low energies in the gas-phase absorption spectrum. To represent such states, we have employed, as previously done,[35,63] a specific atomic-type one-electron basis set of diffuse character placed on the molecular centroid. The lowest Rydberg series will be represented by excitations (basically single excitations) from the HOMO orbital to each of the orbitals of the $n = 3$ series, 3s3p3d, where $n$ has a value one unit more than the valence main quantum number. As the required valence $\pi\pi^*$ active space is small, previous studies at the CASPT2 level[35,63] employed an active space of two electrons in 11 orbitals, including the two valence $\pi\pi^*$ plus the nine 3s3p3d Rydberg orbitals. As was soon detected in polyenes,[63] the CASSCF procedure is unable to deal properly with the simultaneous calculation of valence and Rydberg states. The lack of correlation leads to wave functions in which the MOs are strongly mixed—the so-called valence-Rydberg mixing—yielding too diffuse valence states and too compact Rydberg orbitals that only the multistate CASPT2 is able to correct. Compared with these previous calculations, the RASPT2 results in Table 3 can help us to answer several questions. First, is there any simple partition of the active space that avoids the costly inclusion of the Rydberg orbitals within the CAS space? Second, what is the origin of the valence−Rydberg mixing, and how does RASPT2 handle this problem? Finally, is the multistate treatment still needed, and is there any affordable additional solution?

As observed in Table 3, and in previous studies,[63] for the lowest-energy $^1B_{1u}$ states, the perturbative CASPT2(2,11) correction produces values off by almost 0.5 eV compared to experimental results. The analysis of the orbital extension $\langle r^2 \rangle$ (see that elsewhere)[35] indicates that even when for the lowest-energy state it should reflect its valence and compact character, yielding a similar value to that for the ground state ($1^1A_g$), the magnitude for the orbital extension is almost 4 times larger for both excited states, an illustration of the

***Table 3.*** Excitation Energies (eV) of Selected States of Ethene ($D_{2h}$)[a]

| | | | S | | SDT | | |
|---|---|---|---|---|---|---|---|
| state | CASPT2 (2,11)[b] | MS-CASPT2 (2,11)[b] | RASPT2[c] (2,0,1;0,2,9) | MS-RASPT2[c] (2,0,1;0,2,9) | RASPT2[d] (12,3,3;5,2,16) | MS-RASPT2[d] (12,3,3;5,2,16) | exptl[e] |
| $1^1B_{1u}(\pi\pi^*)$ | 8.43 | 8.04 | 8.44 | 8.13 | 8.06 | 8.00 | 8.0[f] |
| $2^1B_{1u}(3d\pi)$ | 8.98 | 9.38 | 9.03 | 9.35 | 9.35 | 9.41 | 9.33 |

[a] Comparison of CASPT2, MS-CASPT2, RASPT2, and MS-RASPT2 results with $\pi\pi^*$ plus Rydberg and $\pi\pi^*$ and $\sigma\sigma^*$ plus Rydberg active spaces. [b] CASPT2 and MS-CASPT2, from a state average of two $^1B_{1u}$ roots and a two-electron−11-orbital including the two $\pi\pi^*$ MOs and nine ($n = 3$) Rydberg MOs. [c] RASPT2 and MS-RASPT2, from a state average of two $^1B_{1u}$ roots and a two-electron−11-orbital including the two $\pi\pi^*$ MOs (in RAS2) and nine ($n = 3$) Rydberg MOs (in RAS3). Only one particle is allowed (S excitations) in RAS3. [d] RASPT2 and MS-RASPT2, from a state average of three $^1B_{1u}$ roots and a 12-electron−16-orbital including the two $\pi\pi^*$ (in RAS2) MOs, five $\sigma\sigma^*$ MOs (RAS1 and RAS3), and nine ($n = 3$) Rydberg MOs (in RAS3). [e] Experimental data. See ref 7. [f] Estimated vertical excitation energy from earlier theoretical work. See therein, ref 63.

mixed character of the obtained wave function. It was already proven[35] that the use of the MS-CASPT2 level of calculation is required to get a correct result for the interacting $^1B_{1u}$ states and solve the so-called valence−Rydberg mixing problem. After the orthogonalization produced by the MS treatment, the valence and Rydberg states are clearly separated, and the corresponding orbital extension—computed by using the perturbatively-modified CAS-CI (PMCAS-CI) wave function obtained from the MS method—decreases close to the ground state value for the valence $1^1B_{1u}$ state, whereas it largely increases for the Rydberg $2^1B_{1u}$ states.

In the RASPT2 calculations, we have followed two types of computational strategies. First, we have placed the nine Rydberg MOs into the RAS3 active space, leaving the RAS1 space empty and the two $\pi\pi^*$ valence MOs and electrons in RAS2. We have allowed only combined single excitations toward RAS3, since the Rydberg states are typically well described just by single one-electron promotions, as previously suggested.[64] The active space employed can be labeled as RASPT2(2,0,1;0,2,9)(S), including two active electrons and 11 MOs. Table 3 shows that at such a level of calculation, that is, by moving the Rydberg MOs to RAS3, there is no loss of accuracy compared to CASPT2(2,11)/MS-CASPT2(2,11). In this small system, the computational effort is only marginally decreased, but the gain will be much more important in larger molecules. Furthermore, two more advantages can be highlighted: additional valence MOs can eventually be added to RAS2 if required for larger systems, and a single partition of the active space is made available, simplifying considerably the calculations. On the other hand, the behavior of CASPT2(2,11) and RASPT2(2,0,1;0,2,9)(S) with respect to the valence−Rydberg mixing problem is basically the same, as could be expected by the fact that the same types of correlation effects ($\pi\pi^*$ and Rydberg in both cases) are included in the wave function. Still, the MS treatment at the MS-RASPT2 level is required to get correct energies and MO extensions.

RASPT2 allows for enlarging the active space with additional MOs, for instance, the $\sigma$ valence space ($5\sigma$, $5\sigma^*$), and these were incorporated into RAS1 and RAS3 spaces in the calculations labeled RASPT2(12,3,3;5,2,1)(SDT) in Table 3. New states, such as $\sigma\pi^*$, $\pi\sigma^*$, $\sigma\sigma^*$, or $\sigma$Rydberg* can now be described by this method, but not only that. At the RASPT2(12,3,3;5,2,16)(SDT) level, the valence−Rydberg mixing is already solved, and the multistate treatment is not required. As observed, two additional $\sigma^*$ MOs were finally added to the RAS3 space in order to avoid large intruder state problems. Apart from that, the inclusion of the remaining valence electron and orbitals in the active space was sufficient to provide an improved wave function and final results within 0.05 eV from experimental results. This type of behavior has been observed before when active spaces were enlarged to include correlation effects between MOs of different angular moments.[34] Notice that neither the Rydberg nor the $\sigma\sigma^*$ MOs are included in RAS2.

Therefore, in order to incorporate simultaneously the effects of these MOs, a SD excitation level was insufficient (leading to deviations larger than 2 eV) because of the lack of balance between the ground and excited states, as shown in the previous section for FBP. We used one strategy which worked for FBP to balance the treatment; namely, we increased the level of excitation to SDT. Another option would have been to put all the MOs in RAS2, but this would have been unaffordable in this case. We can conclude that RASPT2 provides two different solutions to the valence−Rydberg mixing problem, either reaching the MS level of calculation or introducing new MOs into the RAS spaces, if possible.

Table 4 presents a comparison between the CASSCF/CASPT2/MS-CASPT2(2,11) and RASSCF/RASPT2/MS-RASPT2(2,0,1;0,2,9)(S) levels of calculation for the low-lying singlet and triplet valence and Rydberg states in ethene. As in the previous cases, moving the Rydberg orbitals from RAS2 to RAS3, including up to single excitations from RAS1 to RAS3, provides the same type of accuracy as the full inclusion of the Rydberg MOs into RAS2. This recipe is reliable and a much less costly alternative for the simultaneous calculation of valence and Rydberg states, especially useful for larger systems.

**3.C. Acenes: Benzene and Naphthalene.** In Tables 5, 6, and 7, we report excitation energies for benzene and naphthalene at different levels of theory. In Table 5, valence and Rydberg ($n = 3$ series) singlet excited states of $\pi\pi^*$, $\pi\sigma^*$, and $\sigma\sigma^*$ character calculated with the CASSCF/CASPT2, RASSCF/RASPT2, MS-CASPT2/MS-RASPT2, and CCSD methods are presented. Two RASPT2 strategies have been followed. In the first set of calculations, the six $\pi\pi^*$ valence MOs were left in RAS2, and the nine Rydberg orbitals were placed in RAS3, allowing up to single excitations. As in the case of ethene, no loss of accuracy is observed with respect to CASPT2 when using this procedure, which largely reduces the computational effort. For instance, the active spaces CAS(6,15) and RAS(6,0,1;0,6,9) generate 2345 and 211 CSFs of $^1A_g$ symmetry, respectively (see also

***Table 4.*** Excitation Energies (eV) of the Singlet and Triplet Valence $\pi\pi^*$ and $n = 3$ Rydberg States of Ethene

| | CAS(2,11)[a] | | | RAS(2,0,1;0,2,9)(S)[b] | | | |
|---|---|---|---|---|---|---|---|
| state | CASSCF | CASPT2 | MS-CASPT2 | RASSCF | RASPT2 | MS-RASPT2 | exptl[c] |
| $1^1A_g$ | | | | | | | |
| $1^1B_{3u}$ (3s) | 6.57 | 7.26 | 7.26 | 6.45 | 7.23 | 7.23 | 7.11 |
| $1^1B_{1g}$ (3p$\sigma$) | 7.17 | 7.91 | 7.91 | 7.05 | 7.88 | 7.88 | 7.80 |
| $1^1B_{2g}$ (3p$\sigma$) | 7.18 | 7.91 | 7.91 | 7.06 | 7.89 | 7.89 | 7.90 |
| $1^1B_{1u}$ (V) | 7.93 | 8.43 | 8.04 | 7.83 | 8.44 | 8.13 | 8.0[d] |
| $2^1A_g$ (3p$\pi$) | 7.83 | 8.31 | 8.31 | 7.72 | 8.26 | 8.27 | 8.28 |
| $2^1B_{3u}$ (3d$\sigma$) | 8.01 | 8.81 | 8.81 | 7.88 | 8.78 | 8.78 | 8.62 |
| $3^1B_{3u}$ (3d$\delta$) | 8.11 | 8.93 | 8.93 | 7.98 | 8.90 | 8.90 | 8.90 |
| $1^1B_{2u}$ (3d$\delta$) | 8.11 | 8.96 | 8.96 | 7.98 | 8.94 | 8.94 | 9.05 |
| $1^1A_u$ (3d$\pi$) | 8.10 | 8.93 | 8.93 | 7.97 | 8.91 | 8.91 | |
| $2^1B_{1u}$ (3d$\pi$) | 9.38 | 8.98 | 9.38 | 9.37 | 9.03 | 9.35 | 9.33 |
| $1^3B_{1u}$ (V) | 4.30 | 4.44 | 4.44 | 4.18 | 4.41 | 4.42 | 4.36 |
| $1^3B_{3u}$ (3s) | 6.49 | 7.17 | 7.17 | 6.36 | 7.15 | 7.15 | 6.98 |
| $1^3B_{1g}$ (3p$\sigma$) | 7.14 | 7.87 | 7.87 | 7.02 | 7.86 | 7.86 | 7.79 |
| $1^3B_{2g}$ (3p$\sigma$) | 7.15 | 7.88 | 7.88 | 7.02 | 7.85 | 7.85 | |
| $2^3A_g$ (3p$\pi$) | 7.31 | 8.17 | 8.17 | 7.19 | 8.11 | 8.11 | 8.15 |
| $2^3B_{3u}$ (3d$\sigma$) | 7.99 | 8.79 | 8.80 | 7.86 | 8.77 | 8.78 | 8.57 |
| $3^3B_{3u}$ (3d$\delta$) | 8.05 | 8.88 | 8.89 | 7.92 | 8.86 | 8.86 | |
| $1^3B_{2u}$ (3d$\delta$) | 8.08 | 8.94 | 8.94 | 7.95 | 8.92 | 8.92 | |
| $1^3A_u$ (3d$\pi$) | 8.10 | 8.94 | 8.94 | 7.97 | 8.92 | 8.92 | |
| $2^3B_{1u}$ (3d$\pi$) | 8.41 | 9.09 | 9.10 | 8.28 | 9.04 | 9.04 | |

[a] CASSCF, CASPT2, and MS-CASPT2 results, two electrons and 11 MOs including the two $\pi\pi^*$ MOs and nine ($n = 3$) Rydberg MOs. [b] RASSCF, RASPT2, and MS-RASPT2 results, two electrons and 11 orbitals including the two $\pi\pi^*$ MOs (in RAS2) and nine ($n = 3$) Rydberg MOs (in RAS3). Only one particle is allowed (S excitations) in RAS3. [c] See ref 7. [d] Estimated vertical excitation energy from earlier theoretical work. See therein, ref 63.

***Table 5.*** Excitation Energies (eV) for the Low-Lying Valence and Rydberg Singlet States of Benzene[a]

| | | | S | | SDT | | | |
|---|---|---|---|---|---|---|---|---|
| state | CASPT2[b] | MS-CASPT2[b] | RASPT2[c] (6,0,1;0,6,9) | MS-RASPT2[c,d] (6,0,1;0,6,9) | RASPT2[e] (12,3,3;3,6,12) | MS-RASPT2[d,e] (12,3,3;3,6,12) | CCSD[f] | exptl[g] |
| V-$\pi\pi^*$ | | | | | | | | |
| $1^1B_{2u}$ | 4.94 | 4.94 | 4.98 | 4.98 | 4.72 | 4.93 | 5.19 | 4.90 |
| $1^1B_{1u}$ | 6.22 | 6.21 | 6.20 | 6.20 | 5.83 | 6.44 | 6.59 | 6.20 |
| $1^1E_{1u}$ | 7.12 | 6.92 | 7.00 | 6.82 | 6.72 | 6.93 | 7.17 | 6.94 |
| $2^1E_{2g}$ | 8.05 | 8.05 | 8.09 | 8.10 | 7.93 | 7.94 | 9.18 | 7.8 |
| R-$\pi\pi^*$ | | | | | | | | |
| $2^1E_{1u}$ (3p$\pi$) | 7.22 | 7.29 | 7.28 | 7.30 | 7.16 | 7.39 | 7.58 | 7.41 |
| $2^1A_{1g}$ (3d$\pi$) | 7.88 | 7.87 | 7.94 | 7.93 | 7.85 | 7.87 | 7.86 | 7.81 |
| $1^1E_{2g}$ (3d$\pi$) | 7.91 | 7.88 | 7.95 | 7.96 | 7.88 | 7.85 | 7.85 | 7.81 |
| $1^1A_{2g}$ (3p$\pi$) | 7.89 | 7.91 | 7.93 | 7.92 | 7.82 | 7.84 | 7.88 | |
| R-$\pi\sigma^*$ | | | | | | | | |
| $1^1E_{1g}$ (3s) | 6.54 | 6.54 | 6.54 | 6.54 | 6.50 | 6.50 | 6.55 | 6.33 |
| $1^1A_{2u}$ (3p$\sigma$) | 7.12 | 7.12 | 7.08 | 7.08 | 7.10 | 7.06 | 6.99 | 6.93 |
| $1^1E_{2u}$ (3p$\sigma$) | 7.22 | 7.22 | 7.24 | 7.24 | 7.18 | 7.11 | 7.06 | 6.95 |
| $1^1A_{1u}$ (3p$\sigma$) | 7.15 | 7.15 | 7.16 | 7.16 | 7.11 | 7.18 | 7.14 | |
| $1^1B_{2g}$ (3d$\sigma$) | 7.75 | 7.79 | 7.75 | 7.75 | 7.67 | 7.69 | 7.66 | |
| $1^1B_{1g}$ (3d$\sigma$) | 7.76 | 7.79 | 7.75 | 7.75 | 7.68 | 7.66 | 7.66 | |
| $2^1E_{1g}$ (3d$\delta$) | 7.73 | 7.72 | 7.73 | 7.71 | 7.69 | 7.69 | 7.64 | 7.54 |
| $3^1E_{1g}$ (3d$\delta$) | 7.77 | 7.76 | 7.77 | 7.79 | 7.71 | 7.74 | 7.70 | |
| R-$\sigma\sigma^*$ | | | | | | | | |
| $3^1E_{2g}$ ($\sigma$3s) | | | | | 9.08 | 9.38 | 9.39 | |

[a] For degenerated $D_{6h}$ states, two similar values are obtained in $D_{2h}$ in the CASPT2 and RASPT2 steps, unlike in CASSCF or RASSCF, where external constraints avoid the orbital mixing and symmetry breaking. In all cases, the highest-energy solution has been selected. [b] The CAS space differs for each symmetry (see SI). It includes the six valence $\pi\pi^*$ orbitals and those Rydberg orbitals required to obtain the Rydberg states. All energies referred to ground states with the equivalent CAS. [c] RAS1 empty, RAS2 $\pi\pi^*$ valence MOs, and RAS3 including nine ($n = 3$) Rydberg MOs. A single particle (S) allowed in RAS3. [d] A single-root calculation for the $1^1A_g$ ground state was used in the MS results. [e] Six additional $\sigma\sigma$ electrons and MOs added to RAS1 and RAS3, up to three holes/particles allowed in RAS1/RAS3. [f] Linear response-CCSD calculations.[65] [g] See revision of data in ref 65.

SI). In benzene, because of its high symmetry, the spurious mixing of valence and Rydberg wave functions is not such a problem as it is in ethene, and therefore there is no significant difference when introducing the MS correction, except for symmetries with close-lying valence and Rydberg states like $^1E_{1u}$, where the changes in energies reach up to 0.18 eV. In the second set of calculations, also reported in Table 5, six additional $\sigma\sigma^*$ electrons and MOs have been included, three in RAS1 and three in RAS3, and up to triple excitations have been allowed. These calculations, RASPT2/MS-RASPT2(12,3,3;3,6,12)(SDT), are much more expensive than the previous ones, RASPT2/MS-RASPT2(6,0,1;0,6,9)(S)

***Table 6.*** Excitation Energies (eV) of the Low-Lying Singlet and Triplet Valence $\pi\pi^*$ States of Benzene Optimized at DFT/B3LYP Level Using TZVP Basis Set[a]

| | RASPT2(6,m,m;3,0,3)[b] | | | | | | |
|---|---|---|---|---|---|---|---|
| state | SD | SD/SDT[c] | SDT | SDTQ | CASPT2[d] | RASPT2[e] | exptl[f] |
| $1^1A_{1g}$ | | | | | | | |
| $1^1B_{2u}$ ($\pi\pi^*$) | 5.28 | 5.01 | 5.02 | 5.02 | 5.02 | 4.93 | 4.90 |
| $1^1B_{1u}$ ($\pi\pi^*$) | 6.39 | 6.23 | 6.24 | 6.34 | 6.35 | 6.44 | 6.20 |
| $2^1E_{1u}$ ($\pi\pi^*$) | 7.19 | 6.80 | 6.81 | 6.78 | 6.89 | 6.93 | 6.94 |
| $1^1E_{2g}$ ($\pi\pi^*$) | 8.16 | 7.60 | 7.61 | 7.62 | 7.61 | 7.94 | 7.8 |
| $1^3B_{1u}$ ($\pi\pi^*$) | 4.48 | 4.14 | 4.15 | 4.13 | 4.18 | | 3.94 |
| $1^3E_{1u}$ ($\pi\pi^*$) | 4.96 | 4.79 | 4.80 | 4.85 | 4.85 | | 4.76 |
| $1^3B_{2u}$ ($\pi\pi^*$) | 4.96 | 5.47 | 5.48 | 5.60 | 5.61 | | 5.60 |
| $1^3E_{2g}$ ($\pi\pi^*$) | 7.82 | 7.42 | 7.43 | 7.26 | 7.43 | | 7.24−7.74 |

[a] Leaving the RAS2 space empty is tested in RASPT2. [b] MS-RASPT2(6,m,m;3,0,3)/TZVP, with m being the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. [c] MS-RASPT2(6,2,2;3,0,3)(SD) for the ground state and MS-RASPT2(6,3,3;3,0,3)(SDT) for the excited state. See text. [d] Present MS-CASPT2(6,6)/TZVP, excluding Rydberg orbitals and states. [e] MS-RASPT2(12,3,3;3,6,12)(SDT)/TZVP, including Rydberg orbitals and states. See Table 4. [f] Experimental data. See refs 65 and 66.

(e.g., 927 588 CFSs for $^1A_g$ states), and a similar accuracy is obtained. They slightly improve the results in conflictive states like the $2^1E_{2g}$ valence state, predicted at 7.94 eV at this level, for which the experimental value[65] is 7.8 eV. This highly multiconfigurational state is poorly described by CCSD, which yields 9.18 eV, a value 1.4 eV off with respect to experimental results.[65]

The inclusion of the $\sigma\sigma^*$ MOs and electrons also allows the computation of new states. As an illustration, we have computed the $3^1E_{2g}$ ($\sigma$3s) Rydberg state, a single-reference state, for which RASPT2 and CCSD predict a similar excitation energy.[65] Triple excitations have been included in selected cases in order to compute additional $\sigma\sigma^*$ states. As in the ethene and FBP cases, this is one possible strategy to compensate for the loss of balance caused by not including in RAS2 the $\sigma\sigma^*$ MOs relevant for the simultaneous description of the ground and excited states.

Alternatively, those MOs could be added to RAS2 for both states, and then just up to double excitations would be required.

In Table 6, we compare MS-RASPT2(6,m,m;3,0,3), with $m = 2$ (SD), 3 (SDT), or 4 (SDTQ), to MS-CASPT2(6,6).

Additionally, one set of calculations using SD for the ground and SDT for the excited states has been included. Only the valence $\pi\pi^*$ states are considered. In RASPT2, the RAS2 space was left empty. Inclusion of only up to double excitations leads to errors of about 0.5 eV toward high energies, both in singlet and triplet states. As already shown in FBP, this deviation is due to the lack of balance between the ground and excited states, because the relevant MOs required describing the excited states are excluded from RAS2. To partially correct for this unbalance, we have used the strategies already shown for FBP: either combining SD for the ground state and SDT for the excited states or using SDT or SDTQ for all states. This is a useful comparison between various RASPT2 partitions and the equivalent CASPT2 treatment. In order to reproduce the experimental values, the simultaneous inclusion of Rydberg basis functions, MOs, and states would be required, even to treat valence states only. In conclusion, the strategy of leaving RAS2 empty does not provide extremely accurate results unless a high RAS1/RAS3 excitation level is employed, and it is especially inadequate if only SD is used for all states. Depending on the individual case, it might be preferable either to include in RAS2 the relevant orbitals (see the FBP case) or to increase the excitation level (SDT seems to work for benzene).

Similar comparisons are presented in Table 7 for the singlet and triplet valence $\pi\pi^*$ states of naphthalene. MS-RASPT2 calculations, in which the $\pi\pi^*$ MOs and electrons are placed in RAS1 and RAS3 and RAS2 is left empty, compare reasonably well with the MS-CASPT2(10,10) valence results, but only when at least up to triple (SDT) excitations are considered (at least for the excited states). Otherwise, just by including double excitations (SD), deviations up to 0.8 eV are observed, for instance, for the $2^1B_{3u}$ state (results not included here). The addition of quadruple excitations (SDTQ) has a large effect on the higher-lying singlet states. As for prior cases, we emphasize two aspects: (i) the lack of the relevant MOs in the RAS2 space provides a poor reference description, and (ii) the valence space alone (in the absence of Rydberg MOs) cannot be used to get accurate values with respect to experimental results, except for the lowest-lying states. In Table 7, we also report previous CASPT2(10,11) results in which Rydberg MOs and states were considered.[67] In this case, the CASPT2/RASPT2 method yields its expected level of accuracy, 0.1−0.3 eV.

At this point, we should notice that it is not possible to make a direct comparison between the old calculations and the present ones because, besides the use of Rydberg orbitals in the old calculations, the IPEA shift is not the same (see Computational Details). However, we still report the old results because it is always useful to collect in a single document several results on the same system, and the old results are more directly comparable to experimental results because of the presence of the Rydberg basis functions in the basis set and Rydberg orbitals in the active space. The same is also true for all results reported in Tables 8−17.

**3.D. Heterocyclic Compounds: Furan, Pyrrole, Pyridine, Pyrazine, and Pyrimidine.** As for prior cases, the calculations on these organic heterocyclic molecules have the purpose of establishing the accuracy of the partition in which the RAS2 space is left empty, while the valence $\pi\pi^*$ electrons and MOs are located in RAS1 and RAS3. This partition is very appealing because of its simplicity and its great potential for larger systems, but it requires careful checking. Once again the comparison will be performed toward valence CASPT2 calculations, in which the full valence space ($\pi\pi^*$ and $n$ lone-pair MOs) was included in the CAS. The comparison toward the experimental values would require the simultaneous inclusion of Rydberg MOs and states, as shown previously.[63] Technical details about the calculations are reported in the SI.

An inspection of Tables 8 (furan) and 9 (pyrrole) shows that the RASPT2(6,m,m;3,0,2) calculations increase their accuracy with respect to valence CASPT2 in the order SDT, SD/SDT, and SDTQ (excitation level of RAS1 and RAS3).

***Table 7.*** Excitation Energies (eV) of the Singlet and Triplet Valence $\pi\pi^*$ States of Naphthalene ($D_{2h}$), Leaving the RAS2 Space Empty

| | RASPT2(10,m,m;5,0,5)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_g$ | | | | | | |
| $1^1B_{3u}$ ($\pi\pi^*$) | 4.25 | 4.29 | 4.23 | 4.26 | 4.03 | 3.97, 4.0 |
| $1^1B_{2u}$ ($\pi\pi^*$) | 4.65 | 4.69 | 4.61 | 4.62 | 4.56 | 4.45, 4.7 |
| $2^1A_g$ ($\pi\pi^*$) | 5.98 | 6.02 | 6.00 | 6.05 | 5.39 | 5.50, 5.52 |
| $1^1B_{1g}$ ($\pi\pi^*$) | 5.79 | 5.83 | 5.87 | 5.94 | 5.53 | 5.27, 5.22 |
| $2^1B_{3u}$ ($\pi\pi^*$) | 5.94 | 5.98 | 6.20 | 6.05 | 5.54 | 5.63, 5.55, 5.89 |
| $2^1B_{2u}$ ($\pi\pi^*$) | 6.17 | 6.21 | 6.12 | 6.13 | 5.93 | 6.14, 6.0 |
| $2^1B_{1g}$ ($\pi\pi^*$) | 6.74 | 6.79 | 6.35 | 6.34 | 5.87 | 6.01, 6.05 |
| $3^1A_g$ ($\pi\pi^*$) | 6.77 | 6.81 | 6.66 | 6.72 | 6.04 | |
| $1^3B_{2u}$ ($\pi\pi^*$) | 3.22 | 3.26 | 3.21 | 3.26 | 3.04[f] | 2.98[f,g] |
| $1^3B_{3u}$ ($\pi\pi^*$) | 0.90 | 0.94 | 0.90 | 0.96 | 0.80 | |
| $1^3B_{1g}$ ($\pi\pi^*$) | 1.22 | 1.26 | 1.23 | 1.27 | 1.14 | 1.30 − 1.35[g] |
| $2^3B_{2u}$ ($\pi\pi^*$) | 1.32 | 1.36 | 1.41 | 1.38 | 1.20 | |
| $2^3B_{3u}$ ($\pi\pi^*$) | 1.48 | 1.52 | 1.45 | 1.51 | 1.36 | |
| $1^3A_g$ ($\pi\pi^*$) | 2.22 | 2.26 | 2.25 | 2.28 | 2.18 | 2.25[g] |
| $2^3B_{1g}$ ($\pi\pi^*$) | 2.68 | 2.72 | 2.72 | 2.70 | 2.61 | 3.12[g], 3.0[g] |
| $2^3A_g$ ($\pi\pi^*$) | 3.03 | 3.07 | 3.04 | 3.03 | 2.73 | |
| $3^3A_g$ ($\pi\pi^*$) | 3.15 | 3.19 | 3.14 | 3.17 | 2.81 | 2.93[g] |
| $3^3B_{1g}$ ($\pi\pi^*$) | 3.42 | 3.46 | 3.40 | 3.39 | 3.14 | |

[a] MS-RASPT2(10,m,m;5,0,5)/TZVP results, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. [b] MS-RASPT2(10,2,2;5,0,5)(SD) for the ground state and MS-RASPT2(10,2,2;5,0,5)(SDT) for the excited state. See text. [c] Present MS-CASPT2(10,10)/TZVP results, excluding Rydberg orbitals and states. [d] CASPT2(10,11), ANO-L 3s2p1d/2s+2s2p2d. Rubio et al.,[67] including Rydberg orbitals and states. [e] Experimental optical data in gas phase and solution: George and Morris,[68] Huebner et al.,[69] Mikami and Ito,[70] Dick and Hohlneicher,[71] Klevens and Platt,[72] Bree and Trirunamachandran.[73] [f] Lowest-lying singlet ($1^1A_g$)−triplet ($1^3B_{2u}$) vertical excitation and band absorption maximum. The other excitation energies for the triplet states are referred to the $1^3B_{2u}$ triplet state. [g] Triplet−triplet absorption experimental data: Hunziker[74,75] and Meyer et al.[76]

***Table 8.*** Excitation Energies (eV) of the Low-Lying Singlet and Triplet Valence $\pi\pi^*$ States of Furan ($C_{2v}$), Leaving the RAS2 Space Empty

| | RASPT2(6,m,m;3,0,2)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_1$ | | | | | | |
| $1^1B_2$ ($\pi\pi^*$) | 6.55 | 6.58 | 6.37 | 6.28 | 6.04 | 6.06 |
| $2^1A_1$ ($\pi\pi^*$) | 6.61 | 6.64 | 6.49 | 6.47 | 6.16 | |
| $3^1A_1$ ($\pi\pi^*$) | 8.40 | 8.43 | 8.05 | 8.04 | 7.74 | 7.82 |
| $1^3B_2$ ($\pi\pi^*$) | 4.51 | 4.54 | 4.54 | 4.28 | 3.99 | 4.02 |
| $1^3A_1$ ($\pi\pi^*$) | 5.76 | 5.79 | 5.56 | 5.53 | 5.15 | 5.22 |

[a] MS-RASPT2(6,m,m;3,0,2)/TZVP, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbital and states. [b] MS-RASPT2(6,2,2;3,0,2)(SD) for the ground state and MS-RASPT2(6,3,3;3,0,2)(SDT) for the excited state. See text. [c] Present MS-CASPT2(6,5)/TZVP, excluding Rydberg orbital and states. [d] CASPT2(6,10), ANO-L 4s3p1d/2s1p+2s2p2d. Serrano-Andrés et al.[63] including Rydberg orbitals and states. [e] Experimental data. Flicker et al.[78]

***Table 9.*** Excitation Energies (eV) of the Low-Lying Singlet and Triplet Valence $\pi\pi^*$ States of Pyrrole ($C_{2v}$), Leaving the RAS2 Space Empty

| | RASPT2(6,m,m;3,0,2)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_1$ | | | | | | |
| $2^1A_1$ ($\pi\pi^*$) | 6.40 | 6.54 | 6.30 | 6.28 | 5.92 | |
| $1^1B_2$ ($\pi\pi^*$) | 6.82 | 6.96 | 6.67 | 6.62 | 6.00 | 5.98 |
| $3^1A_1$ ($\pi\pi^*$) | 8.18 | 8.32 | 7.94 | 7.92 | 7.46 | 7.54 |
| $1^3B_2$ ($\pi\pi^*$) | 4.64 | 4.78 | 4.50 | 4.48 | 4.27 | 4.21 |
| $1^3A_1$ ($\pi\pi^*$) | 5.22 | 5.35 | 5.29 | 5.43 | 5.16 | 5.10 |

[a] MS-RASPT2(6,m,m;3,0,2)/TZVP, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbital and states. [b] MS-RASPT2(6,2,2;3,0,2)(SD) for the ground state and MS-RASPT2(6,3,3;3,0,2)(SDT) for the excited state. See text. [c] Present MS-CASPT2(6,5)/TZVP, excluding Rydberg orbital and states. [d] CASPT2(6,10), ANO-L 4s3p1d/2s1p+2s2p2d,[63] including Rydberg orbitals and states. [e] Experimental data. Flicker et al.,[78] Bavia et al.,[79] and Van Veen.[80]

The RAS(SDTQ) partition typically yields results very close to the full CAS calculation.[77]

Only when reaching a SDTQ level of excitation, which can be prohibitive for larger molecules, can the calculations be considered really accurate. This shows that in some cases it might not be practical to leave RAS2 empty when computing excited states. In order to compare with experimental results, the Rydberg MOs and states may have to be included in the calculation, as shown by previous CASPT2 calculations,[63] whose accuracy was established to be within 0.1 eV.

Tables 10, 11, and 12 provide data for the same type of calculations for the azabenzenes pyridine, pyrazine, and pyrimidine. Both valence $\pi\pi^*$ and $n\pi^*$ states are considered. The performance of RASPT2(10,m,m;5,0,3) versus CASPT2(10,8) is similar to that in the furan and pyrrole cases, although in pyridine the $n\pi^*$ states are less accurately described at the SDT level than at the SD/SDT level. The effect is less pronounced for the other two molecules. In general, we observe that for states below 7.0 eV the deviation of the SDT and SDTQ RAS calculations falls within a value of 0.2 eV compared to valence CASPT2. On the other hand, for higher lying states, the deviation can reach up to 0.5 and 0.3 eV at the SDT and SDTQ levels, respectively. It can be therefore concluded that the empty-RAS2 approach should be used with caution. Even for low-lying roots, a SD or SDT level of excitation may not suffice to obtain a 0.2 eV accuracy, and combining the SD and SDT levels could be a better alternative (considering that extending to SDTQ is

***Table 10.*** Excitation Energies (eV) of the Singlet and Triplet Valence $\pi\pi^*$ and $n\pi^*$ States of Pyridine ($C_{2v}$)

| | RASPT2(8,m,m;4,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_1$ | | | | | | |
| $1^1B_1$ ($n\pi^*$) | 5.07 | 5.40 | 5.15 | 5.05 | 4.91 | 4.59 |
| $1^1A_2$ ($n\pi^*$) | 5.38 | 5.71 | 5.42 | 5.35 | 5.17 | 5.43 |
| $1^1B_2$ ($\pi\pi^*$) | 4.74 | 5.03 | 5.26 | 5.10 | 4.84 | 4.99 |
| $2\ ^1A_1$ ($\pi\pi^*$) | 6.47 | 6.68 | 6.71 | 6.57 | 6.42 | 6.38 |
| $3^1A_1$ ($\pi\pi^*$) | 7.31 | 7.49 | 7.48 | 7.12 | 7.23 | 7.22 |
| $2^1B_2$ ($\pi\pi^*$) | 7.16 | 7.10 | 7.31 | 7.17 | 7.48 | |
| $4^1A_1$ ($\pi\pi^*$) | 8.25 | 8.51 | 8.47 | 8.23 | 7.96 | |
| $3^1B_2$ ($\pi\pi^*$) | 8.07 | 8.34 | 8.42 | 8.21 | 7.94 | |
| $1^3A_1$ ($\pi\pi^*$) | 4.34 | 4.67 | 4.44 | 4.32 | 4.05 | 4.10 |
| $1^3B_1$($n\pi^*$) | 4.52 | 4.84 | 4.69 | 4.50 | 4.41 | |
| $1^3B_2$ ($\pi\pi^*$) | 4.80 | 5.09 | 4.89 | 4.82 | 4.56 | 4.84 |
| $2^3A_1$ ($\pi\pi^*$) | 5.00 | 5.34 | 5.18 | 5.02 | 4.73 | |
| $1^3A_2$ ($n\pi^*$) | 5.37 | 5.71 | 5.42 | 5.36 | 5.10 | |
| $2^3B_2$ ($\pi\pi^*$) | 6.40 | 6.74 | 6.64 | 6.69 | 6.02 | |
| $3^3A_1$ ($\pi\pi^*$) | 7.71 | 8.04 | 7.86 | 7.68 | 7.34 | |
| $3^3B_2$ ($\pi\pi^*$) | 7.17 | 7.53 | 6.97 | 6.88 | 7.28 | |

[a] MS-RASPT2(8,m,m;4,0,3)/TZVP, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. [b] MS-RASPT2(8,2,2;4,0,3)(SD) for the ground state and MS-RASPT2(8,3,3;4,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(8,7)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(8,12), ANO-L 4s3p1d/2s1p. Lorentzon et al.[81] including Rydberg orbitals and states. [e] Experimental data, Bolovinos et al.[82]

***Table 11.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Pyrazine ($D_{2h}$)

| | RASPT2 (10,m,m;5,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_g$ | | | | | | |
| $1^1B_{1u}$ ($n\pi^*$) | 4.06 | 4.21 | 3.95 | 4.09 | 3.85 | 3.83 |
| $1^1A_u$ ($n\pi^*$) | 4.67 | 4.82 | 4.65 | 4.67 | 4.63 | |
| $1^1B_{2u}$ ($\pi\pi^*$) | 5.02 | 5.18 | 4.98 | 5.04 | 4.76 | 4.81 |
| $1^1B_{2g}$ ($n\pi^*$) | 5.55 | 5.71 | 5.39 | 5.55 | | 5.46 |
| $1^1B_{3g}$ ($n\pi^*$) | 6.50 | 6.65 | 6.31 | 6.47 | | 6.10 |
| $1^1B_{3u}$ ($\pi\pi^*$) | 6.61 | 6.77 | 6.43 | 6.68 | 6.69 | 6.51 |
| $2^1B_{3u}$ ($\pi\pi^*$) | 7.82 | 7.97 | 7.46 | 7.57 | 7.53 | 7.67 |
| $1^1B_{2u}$ ($\pi\pi^*$) | 7.55 | 7.70 | 7.51 | 7.44 | 7.74 | 7.67 |
| $1^1B_{1g}$ ($\pi\pi^*$) | 8.33 | 8.48 | 8.37 | 8.43 | 8.31 | |
| $2^1A_g$ ($\pi\pi^*$) | 8.71 | 8.87 | 8.67 | 8.68 | 8.22 | |

[a] MS-RASPT2(10,m,m;5,0,3)/TZVP, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbital and states. [b] MS-RASPT2(10,2,2;5,0,3)(SD) for the ground state and MS-RASPT2(10,3,3;5,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(10,8)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(10,12), ANO-L 4s3p2d/3s2p. Fülscher and Roos[83] including Rydberg orbitals and states. [e] Experimental data. Innes et al.,[84] Bolovinos et al.,[85] and Okuzawa et al.[86]

actually quite expensive). Otherwise, for higher-lying states, the inclusion of Rydberg orbitals is indispensable.

**3.E. DNA/RNA Nucleobases: Adenine, Thymine, Uracil, and Cytosine.** In this section, we describe the results of the CASPT2 and RASPT2 study of the DNA/RNA nucleobases adenine, thymine, uracil, and cytosine. We have employed partitions of the RAS spaces similar to those described in the previous section; namely, the valence $\pi\pi^*$ and lone-pair orbitals have been distributed in RAS1 (Hartree−Fock occupied MOs) and RAS3 (Hartree−Fock unoccupied MOs), and RAS2 has been left empty. The results

***Table 12.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Pyrimidine ($C_{2v}$)

| | RASPT2 (10,m,m;5,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A_1$ | | | | | | |
| $1^1B_1$ ($n\pi^*$) | 4.13 | 4.49 | 4.14 | 4.33 | 3.81 | 3.8 − 4.1 |
| $1^1A_2$ ($n\pi^*$) | 4.58 | 4.94 | 4.55 | 4.71 | 4.12 | 4.62 |
| $1^1B_2$ ($\pi\pi^*$) | 5.23 | 5.35 | 5.45 | 5.33 | 4.23 | 5.12 |
| $2^1A_1$ ($\pi\pi^*$) | 6.67 | 7.04 | 6.96 | 6.96 | 6.7 | 6.7 |
| $3^1A_1$ ($\pi\pi^*$) | 7.71 | 8.08 | 7.75 | 7.54 | 7.57 | 7.57 |
| $2^1B_2$ ($\pi\pi^*$) | 7.50 | 7.60 | 7.57 | 7.37 | 7.32 | 7.57 |
| $4^1A_1$ ($\pi\pi^*$) | 7.86 | 8.22 | 7.82 | 7.86 | 7.82 | |
| $3^1B_2$ ($\pi\pi^*$) | 8.61 | 9.00 | 8.80 | 8.76 | 8.31 | 8.8 |

[a] MS-RASPT2(10,m,m;5,0,3)/TZVP, with m the indicated level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbital and states. [b] MS-RASPT2(10,2,2;5,0,3)(SD) for the ground state and MS-RASPT2(10,3,3;5,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(10,8)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(8,12), ANO-L 4s3p2d/3s2p. Fülscher et al.[87] including Rydberg orbitals and states. [e] Experimental data, Bolovinos et al.[82] See also ref 88.

***Table 13.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Adenine ($C_s$)

| | RASPT2 (12,m,m;6,0,4)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A'$ | | | | | | |
| $2^1A'$ ($\pi\pi^*$) | 5.10 | 5.18 | 5.14 | 5.10 | 5.13 | 4.6 |
| $3^1A'$ ($\pi\pi^*$) | 5.13 | 5.21 | 5.17 | 5.17 | 5.20 | 4.8 − 4.9 |
| $1^1A''$ ($n\pi^*$) | 5.07 | 5.15 | 4.97 | 5.15 | *f* | 5.4 |
| $4^1A'$ ($\pi\pi^*$) | 6.34 | 6.43 | 6.42 | 6.41 | 6.24 | 5.9 − 6.0 |
| $2^1A''$ ($n\pi^*$) | 5.76 | 5.84 | 5.78 | 5.85 | 6.15 | |
| $5^1A'$ ($\pi\pi^*$) | 6.40 | 6.48 | 6.47 | 6.48 | 6.72 | 6.3 − 6.4 |
| $6^1A'$ ($\pi\pi^*$) | 6.57 | 6.65 | 6.63 | 6.65 | 6.99 | 6.8 |

[a] MS-RASPT2(12,m,m;6,0,4)/TZVP, m indicates level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. RAS2 is empty here. [b] MS-RASPT2(12,2,2;6,0,4)(SD) for the ground state and MS-RASPT2(12,3,3;6,0,4)(SDT) for the excited state. See text. [c] Present MS-CASPT2(12,10)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(12,11), ANO-L 4s3p1d/2s1p,[10] including Rydberg orbitals and states. [e] Experimental absorption data in solution. Mixture with the 7H-adenine tautomer has been noticed. See ref 10 for a critical revision of the experimental results. [f] The active spaces lacked the lowest-lying lone-pair orbital, missing therefore the lowest-lying $\pi^*$ state.

are compared with full $\pi\pi^*$ and lone-pair valence and valence plus Rydberg CASPT2 calculations and with experimental data.

Table 13 describes the results for the singlet states of adenine. There is an overall agreement within 0.2−0.3 eV among the various sets of calculations. We have not reported here calculations of the type RASPT2(12,2,2;6,0,4)(SD) (RAS2 space empty), which display deviations close to 0.6 eV for some states. It was shown previously that the SD level of excitation is unreliable for excited states if the RAS2 space is not balanced, including, for the states under consideration, the MOs largely modifying their occupation number in the excitation process. The reader must be warned about the comparison of some of the previous CASPT2 results for $n\pi^*$ transitions. In some cases, the proper lone-

***Table 14.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Thymine ($C_s$)

| | RASPT2(12,m,m;6,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A'$ | | | | | | |
| $1^1A''$ ($n\pi^*$) | 5.01 | 5.29 | 5.09 | 5.24 | 4.77[f] | |
| $2^1A'$ ($\pi\pi^*$) | 5.64 | 5.93 | 5.80 | 5.56 | 4.88 | 4.8 − 5.1 |
| $3^1A'$ ($\pi\pi^*$) | 6.61 | 6.90 | 6.72 | 6.54 | 5.88 | 6.0 − 6.1 |
| $2^1A''$ ($n\pi^*$) | 6.37 | 6.65 | 6.43 | 6.54 | | |
| $4^1A'$ ($\pi\pi^*$) | 6.70 | 6.98 | 6.73 | 6.58 | 6.10 | 6.5 − 6.6 |
| $5^1A'$ ($\pi\pi^*$) | 7.59 | 7.87 | 7.38 | 7.19 | 7.13 | 6.9 − 7.0 |

[a] MS-RASPT2(12,m,m;6,0,3)/TZVP, m indicates the level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. RAS2 is empty here. [b] MS-RASPT2(12,2,2;6,0,3)(SD) for the ground state and MS-RASPT2(12,3,3;6,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(12,9)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(12,11), ANO-L 4s3p1d/2s. Lorentzon et al.,[89] including Rydberg orbitals and tates. [e] Experimental absorption data in gas phase and solution. See ref 89 for a critical revision of the experimental results. [f] See ref 90.

***Table 15.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Uracil ($C_s$)

| | RASPT2(12,m,m;6,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A'$ | | | | | | |
| $1^1A''$ ($n\pi^*$) | 5.19 | 5.36 | 5.56 | 5.43 | 4.80 | |
| $2^1A'$ ($\pi\pi^*$) | 5.84 | 6.01 | 5.93 | 5.90 | 5.00 | 4.8 − 5.1 |
| $3^1A'$ ($\pi\pi^*$) | 6.64 | 6.82 | 6.65 | 6.73 | 5.82 | 6.0 − 6.1 |
| $2^1A''$ ($n\pi^*$) | 6.63 | 6.80 | 6.90 | 6.78 | 6.20 | |
| $4^1A'$ ($\pi\pi^*$) | 6.79 | 6.96 | 6.87 | 6.93 | 6.46 | 6.5 − 6.6 |
| $5^1A'$ ($\pi\pi^*$) | 7.58 | 7.75 | 7.38 | 7.48 | 7.00 | 6.9 − 7.0 |

[a] MS-RASPT2(12,m,m;6,0,3)/TZVP, m indicates the level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. RAS2 is empty here. [b] MS-RASPT2(12,2,2;6,0,3)(SD) for the ground state and MS-RASPT2(12,3,3;6,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(12,9)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(12,11), ANO-L 4s3p1d/2s. Lorentzon et al.,[89] including Rydberg orbitals and sates. [e] Experimental absorption data in gas phase and solution. See ref 89 for a critical revision of the experimental results.

***Table 16.*** Excitation Energies (eV) of the Singlet Valence $\pi\pi^*$ and $n\pi^*$ States of Cytosine ($C_s$)

| | RASPT2 (12,m,m;6,0,3)[a] | | | | | |
|---|---|---|---|---|---|---|
| state | SD/SDT[b] | SDT | SDTQ | CASPT2[c] | CASPT2[d] | exptl[e] |
| $1^1A'$ | | | | | | |
| $2^1A'$ ($\pi\pi^*$) | 4.53 | 4.96 | 5.04 | 4.72 | 4.39 | 4.4 − 4.6 |
| $1^1A''$ ($n_O\pi^*$) | 5.49 | 5.92 | 5.85 | 5.52 | 5.00 | |
| $2^1A''$ ($n_N\pi^*$) | 5.72 | 6.15 | 6.03 | 5.73 | 5.06[f] | |
| $3^1A'$ ($\pi\pi^*$) | 5.79 | 6.22 | 6.29 | 5.95 | 5.36 | 5.0 − 5.5 |
| $4^1A'$ ($\pi\pi^*$) | 6.75 | 7.17 | 7.30 | 6.85 | 6.16 | 5.8 − 6.3 |
| $5^1A'$ ($\pi\pi^*$) | 7.01 | 7.43 | 7.37 | 7.00 | 6.74 | 6.7 − 7.1 |

[a] MS-RASPT2(10,m,m;6,0,3)/TZVP; m indicates the level of excitation: SD, SDT, or SDTQ, excluding Rydberg orbitals and states. RAS2 is empty here. [b] MS-RASPT2(12,2,2;6,0,3)(SD) for the ground state and MS-RASPT2(12,3,3;6,0,3)(SDT) for the excited state. See text. [c] Present MS-CASPT2(12/9)/TZVP, excluding Rydberg orbitals and states. [d] CASPT2(10,12), ANO-L 4s3p1d/2s.[91] including Rydberg orbitals and states. [e] Experimental absorption data in gas phase and solution. See ref 91 for a critical revision of the experimental results. [f] See ref 3. The corresponding lone pair is missing in the active space in ref 91.

pair MO was left outside the active space, and the corresponding state was therefore missing. This is not due to an inaccuracy of the CASPT2 method but instead to a bad selection of the active space.

***Table 17.*** Comparison between CASPT2 and RASPT2 Excitation Energies (eV) of the Excited States of the Nickel Atom[a]

| | CASPT2 | CASPT2 | RASPT2(10,0,m;0,6,5) (3s3p)3d4s4d[b] | | | |
|---|---|---|---|---|---|---|
| States | 3d4s[b] | 3d4s4d[b] | SD | SDT | SDTQ | exptl[c] |
| $^3D$ ($3d^94s^1$) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $^3F$ ($3d^84s^2$) | −0.33 | −0.10 | 0.01 | −0.03 | 0.04 | 0.03 |
| $^1D$ ($3d^94s^1$) | 0.01 | 0.28 | 0.29 | 0.24 | 0.24 | 0.33 |
| $^1D$ ($3d^84s^2$) | 1.16 | 1.45 | 1.52 | 1.52 | 1.61 | 1.59 |
| $^1S$ ($3d^{10}$) | −0.97 | 1.79 | 2.16 | 2.21 | 2.01 | 1.74 |
| $^3P$ ($3d^84s^2$) | 1.48 | 1.68 | 1.79 | 1.76 | 1.85 | 1.86 |
| $^1G$ ($3d^84s^2$) | 2.33 | 2.54 | 2.63 | 2.62 | 2.71 | 2.65 |

[a] The RAS partition is 3d4s in RAS2 and 4d in RAS3, with RAS1 empty, RAS(10,0,m;0,6,5), and different excitation levels in RAS3. Core−valence correlation is computed at the perturbative level (3s3p electrons). [b] CASPT2 and RASPT2(10,0,m;0,6,5) with 18 electrons correlated, basis set ANO-RCC 7s6p4d3f2g, SA(15) for triplet states and SA(19) for singlet states except $^1S(3d^{10})$, a single root calculation. [c] Experimental data. NIST (national institute of standards and technology).[94]

The results on thymine and uracil in Tables 14 and 15 show different trends. In both cases, the RASPT2(SD) level with RAS2-empty was insufficient to get quantitative results. The SDTQ level of excitation is indispensable, which is unfeasible for larger systems. Alternatively, it is better to balance the calculations using SD for the ground state and SDT for the excited states. For higher-lying states, for instance, in the $2^1A'$ and $3^1A'$ $\pi\pi^*$ states, large discrepancies are found between the valence−Rydberg CASPT2 results and the experimental values. The effect is even more clear for both the $\pi\pi^*$ and $n\pi^*$ states of uracil, which is related to the absence of the Rydberg MOs from the active space. The advice would be to add the Rydberg MOs and avoid the MS procedure if the diffuse MOs are not included, because it might lead to overestimated interactions, as was proved previously.[34]

Table 16 displays the results on cytosine, and the conclusions are similar to those obtained for the previous pyrimidine nucleobases. In this case, the SD(ground state)/SDT(excited state) strategy becomes particularly accurate as compared to that with full CASPT2. This strategy could be a cheap alternative in cases where larger RAS2 spaces or high excitation (SDTQ) levels are unfeasible. Once again, one should emphasize the need to include Rydberg MOs and states for calculations on high-lying excited states. For low-lying states, the Rydberg MOs may be excluded, but then the use of the multistate approach is not recommended, because it might lead to spurious interactions.

**3.F. Transition Metal Compounds and the Double d-Shell Effect: The Nickel Atom and the Copper Tetrachloride Dianion.** Because of the strong correlation effects associated with the 3d shell in first-row transition metals (TM), the inclusion of a second correlating d shell (4d) in the active space was shown to have crucial effects on the relative state energies obtained from CASPT2 for molecules containing first-row transition metal atoms with

a more than a half-filled 3d shell.[33,36,43,92,93] This effect, referred to as the double d-shell effect, is manifested in particular when dealing with transitions between states with a different 3d occupation number, e.g., 3d→4s transitions or charge-transfer (CT) transitions. The double d-shell effect was first reported in a CASPT2 study of the low-lying states of the nickel atom.[36] Here, we report the results of a comparative CASPT2/RASPT2 study of the lowest states in the electronic spectra of the nickel atom and the copper tetrachloride dianion ($CuCl_4^{2-}$). The underlying motivation of this study is to check whether it might be possible to treat the double d-shell effect by means of the much cheaper RASPT2 strategy by, for instance, moving the 4d shell into RAS3. This would allow for applicability of the present multiconfigurational approach to more extended and complex TM systems that have so far been inaccessible or could only be treated qualitatively, because of size limitations of the CASSCF active space, e.g., systems with multiple TM centers.[95]

The calculated results obtained for the spectrum of the Ni atom are presented in Table 17 and compared to experimental results. The first two columns show the CASPT2 relative energies obtained with an active space containing 10 electrons in either the minimum valence active space (3d, 4s) or extended with an extra d shell (3d, 4s, 4d). The double d-shell effect is clearly illustrated by these results. The CASPT2 excitation energies obtained without a second d shell in the active space strongly deviate from the experimental data by 0.3−0.5 eV for all states except $^1S$ ($3d^{10}$), for which an exceptionally large error of as much as 2.7 eV is found. After including the second d-shell, the errors are reduced to 0.2 eV for all calculated states. These results might be further improved by also including the (4p) shell into the active space, and by further extending the basis set.

The next three columns in Table 17 give the results obtained from RASPT2(10,0,m;0,6,5), with m being the electrons allowed in RAS3, representing a maximum excitation level from two (SD) up to four (SDTQ). Because of the poor convergence of the RASSCF orbital optimization, the RASSCF(SDT) and RASSCF(SDTQ) energies have been calculated at the CI level without orbital optimization, and using the molecular orbitals converged at the RASSCF(SD) level. As one can see, even at the SD level, the double-shell effect is described reasonably well for most states, the results deviating by at most 0.1 eV with respect to the full CASPT2 results. Minor oscillations are observed when increasing the excitation level to SDT and further to SDTQ, but in general there is no clear sign of a systematic improvement. An exception is again the $^1S$ ($3d^{10}$) state. Here, going from CASPT2 to RASPT2 leads to a significant deterioration of the results, by 0.37 eV at the SD level, and decreasing to 0.22 eV at the SDTQ level. However, it is clear that for this state the RASPT2 description of the double d-shell effect is not converged with respect to the excitation level, and higher levels of excitations are necessary for obtaining quantitative accuracy.

As a final set of calculations, we include here the study at the CASPT2 and RASPT2 levels of the excitation energies of the ligand field (LF) states and a charge transfer (CT) state in the electronic spectrum of the copper tetrachloride dianion ($CuCl_4^{2-}$). In order to compare the results with previous reports,[57,96] the same geometry (planar, $D_{4h}$, with the Cl ligands on the *x* and *y* axes) and basis sets were used. The valence electronic structure of this molecule is presented in Figure 2. The ground state (GS), $1^2B_{1g}$, has a singly occupied molecular orbital (SOMO), $\sigma$-antibonding with predominant Cu $3d_{x^2-y^2}$ character, and the lowest part in the spectrum is built from excitations of an electron out of each of the other four 3d orbitals, giving rise to three ligand field (LF) states $1^2B_{2g}$, $1^2E_g$, and $1^2A_{1g}$. An important charge-transfer (CT) state, $2^2B_{1g}$, corresponding to an excitation out of the bonding counterpart of the ground state SOMO is also included in the calculations. This CT state belongs to the same symmetry representation as the ground state, and it was shown previously[40] that the interaction between both states resulting from a MS-CASPT2 treatment gives rise to a strongly enhanced covalent character of the GS Cu−Cl $\sigma$ bonds, by increasing the chlorine $2p\sigma$ contribution in the GS $b_{1g}$* SOMO. The purpose of the present study is therefore not only to investigate whether the electronic spectrum of $CuCl_4^{2-}$ may be satisfactorily reproduced by means of a RASPT2 rather than a CASPT2 treatment but also to see whether the same covalency enhancing effect for the GS may be obtained from a MS-RASPT2 treatment. The latter may be evaluated by comparing the Mulliken spin populations from the CASSCF and perturbed modified (PM) CASSCF GS wave functions obtained before and after the multistate treatment, respectively.

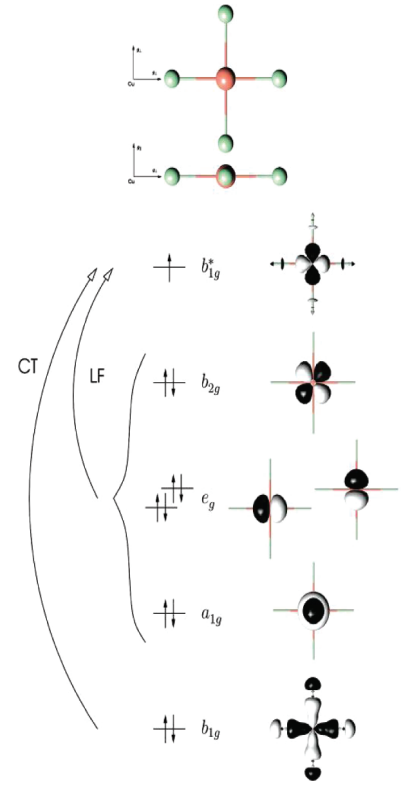



***Figure 2.*** Schematic representation of the geometry and electronic structure of $[CuCl_4]^{2-}$.

The CASPT2 calculations are based on an active space of 11 orbitals, consisting of the Cu 3d and 4d shells together with the bonding $b_{1g}$ orbital. In the RASPT2 calculations,

***Table 18.*** Excitation Energies (eV) of $CuCl_4^{2-}$ Computed at the CASPT2(11,11) and RASPT2(11,0,n;0,6,5) Levels of Calculation Compared with the Available Experimental Data

| states | SS-CASPT2 (11,11) | MS-CASPT2 (11,11) | SS-RASPT2(11,0,m;0,6,5) SD | SDT | SDTQ | MS-RASPT2(11,0,m;0,6,5) SD | SDT | SDTQ | exptl[a] |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $1^2B_{1g}$ (GS) | | | | | |
| $1^2B_{2g}$ (LF) | 1.52 | 1.65 | 1.36 | 1.48 | 1.52 | 1.51 | 1.63 | 1.67 | 1.55 |
| $1^2E_g$ (LF) | 1.77 | 1.90 | 1.60 | 1.72 | 1.76 | 1.75 | 1.87 | 1.91 | 1.76 |
| $1^2A_{1g}$ (LF) | 2.00 | 2.13 | 1.91 | 1.92 | 1.99 | 2.06 | 2.07 | 2.14 | |
| $2^2B_{1g}$ (CT) | 4.60 | 4.86 | 4.51 | 4.52 | 4.42 | 4.81 | 4.81 | 4.73 | |

[a] See refs 57, 96.

the correlating 4d shell was transferred into RAS3, leaving RAS1 empty and the other six orbitals in RAS2. This then gives results of the type RASPT2(11,0,m;0,6,5), with m representing the RAS2→RAS3 excitation level. The calculated excitation energies obtained from either a single-state (SS) or multistate (MS) treatment are presented in Table 18. Looking at the SS results first, we note that for the LF states, the results obtained from RASPT2-SDTQ calculations are virtually indistinguishable from CASPT2. A deterioration of the results is observed when decreasing the RASSCF excitation level to SDT and further to SD, although the accuracy of the results obtained from the latter treatment, within 0.2 eV, is still acceptable. On the other hand, for the CT states, the RASPT2 treatment seems to be more problematic, giving an excitation energy that deviates more from the CASPT2 results as the level of excitation is increased. Only two of the states included in the calculations belong to the same $B_{1g}$ representation. A MS treatment will therefore leave the total energy of the other states unaffected, while stabilizing the $1^2B_{1g}$ ground state and destabilizing the $2^2B_{1g}$ CT state. This then gives rise to a calculated MS-CASPT2 spectrum in which all three LF states are raised in energy by the same amount, 0.13 eV, as compared to SS-CASPT2, while the $2^2B_{1g}$ CT state is raised by twice this amount. The results obtained from MS-RASPT2 follow the same trend with respect to SS-RASPT2. As such, the same conclusions concerning the accuracy obtained from RASPT2 for the LF and CT states may be drawn from Table 18, as already noted for the SS results. As compared to the experimental excitation energies for the $1^2B_{2g}$ and $1^2E_g$ (LF) states, the SS treatment yields better excitation energies than MS-CASPT2. The addition of the MS step does not increase the accuracy of the results at any of the levels, CASPT2 or RASPT2. This is not unexpected because the active space requirements with MS are larger than for the lower-level methods. It has been shown before that the addition of angular correlation, that is, the inclusion of orbitals with different angular momentum quantum numbers in the active space, largely improves the MS results.[34]

It should finally be mentioned that the RASSCF calculations also reproduce the CASSCF Mulliken spin populations for all states (see SI). In particular, for the ground state, the spin population on copper obtained from RASSCF, 0.84, reflects a very ionic Cu−Cl bond. As was shown in a previous study,[57] this ionic description gives rise to calculated EPR g factors that deviate considerably more from the free-electron value than is observed from experimental results. A significant improvement of the calculated g factors may be obtained by making use instead of the PM CASSCF wave function, giving rise to a more covalent description of the Cu−Cl bonds, with a Mulliken spin population on copper that is decreased by 7%, thus approaching the value of 0.62 ± 0.02 deduced from experimental results.

The most important conclusion to be drawn from the results obtained in this section is that, in general, moving the 4d shell into the RAS3 space is a good strategy that leads to much less expensive calculations in transition metal systems without a considerable loss in accuracy. This then allows for the extension of the methodology to larger systems, both increasing the number of transition metal atoms or including additional ligand molecules. For instance, the number of CSFs decreases from near 98 000 in a CASSCF(10,11) calculation to 4300, 19 000, and 47 500 at the RASSCF(SD), SDT, and SDTQ levels, respectively (for the Ni calculations, see SI).

## 4. Summary and Conclusions

The electronic excited states of a number of organic (free base porphin, ethene, benzene, naphthalene, furan, pyrrole, and several azobenzenes and nucleobases) and inorganic (the nickel atom and the copper tetrachloride dianion) systems have been computed at the RASSCF/RASPT2/MS-RASPT2 (RAS) levels of calculation using different active space partitions and strategies. The results have been compared to those obtained with well-established procedures like CASSCF/CASPT2/MS-CASPT2 (CAS) or CCSD, and to experimental values, in order to determine the accuracy of several procedures used to divide the RAS space. Our main goal was to establish computational strategies that would provide the most accurate results at reasonable computational costs that one could eventually employ for larger systems. The RAS approaches have many possible ways to define the active spaces for the multiconfigurational calculation, and therefore systematic selection procedures have to be developed and calibrated.

Free base porphin has been first investigated with several partition procedures. RASPT2 has proved to be an excellent strategy to avoid arbitrary divisions of the $\pi$ space in a system in which the full-$\pi$ active space (26 electrons in 24 MOs) is out of reach for the CASPT2 method. It has been shown that in the RASPT2 method the proper definition of the RAS2 space (in which a full-CI is performed to define the configurational reference space) is crucial to assessing the accuracy of the calculations. In particular, an initial analysis of the occupation numbers displayed by the relevant MOs, even at a simple RASSCF(SD) level of calculation, is very useful to determine the composition of RAS2. When

computing a RASPT2 energy difference, the highest accuracy is obtained when the MOs changing their occupation number from one state (typically the ground state) to the other (an excited state) the most are simultaneously included in the RAS2 space, leaving the other less significant MOs in the RAS1/RAS3 spaces. If this requirement is fulfilled, a single−double (SD) level of excitations in these two latter active spaces partitions is sufficient to get a high accuracy. In free base porphin, as is typical in many other $\pi$ organic systems, Gouterman's four MOs (HOMO, HOMO−1, LUMO, and LUMO+1) form the basic set required to describe the four low-lying $\pi\pi^*$ excited states, and therefore it will be sufficient to include them in RAS2 while leaving the remaining $\pi\pi^*$ MOs in RAS1/RAS3 and reaching a SD level of excitation to get accurate results. Higher states will however require extension of the RAS2 space to include additional MOs. It is possible to design a less straightforward strategy and perform calculations for each of the two states with different active spaces. If the proper MOs are excluded from RAS2, the results are unbalanced in the CI treatment, and the second-order perturbation correction may not be able to compensate the results. Particularly for this case, a SD level of excitation is clearly insufficient. Although not as accurate as the inclusion of the proper MOs in RAS2, there are some additional strategies that may help to slightly improve the results even if some important MOs are excluded from RAS2, for instance, using different levels of excitation for the two considered states, like SD for the ground and SDT for the excited state, or increasing the overall excitation as much as possible, SDT, or even better, SDTQ, although these latter strategies might be impossible to apply because of the very large configurational spaces. All of these results open the possibility to use RASPT2 for many organic systems with extended $\pi$ spaces without a further loss of accuracy due to restrictions in the size of the active space.

Calculations on the valence and Rydberg singlet and triplet excited states of ethene and benzene have illustrated the advantages of RASPT2 versus CASPT2 when large active spaces including both valence and Rydberg states and MOs are required. A new strategy for the active MO partition has been used in which the Rydberg MOs—typically nine ($n$ = 3) for common organic systems—are placed in RAS3, leaving in the RAS2 space the valence $\pi\pi^*$ MOs and electrons, and allowing, apart from the full-CI expansion within RAS2, just single excitations toward RAS3. The advantage of the RAS approach, whose accuracy is similar to that of a full CAS calculation, is that the Rydberg orbitals can be moved out of the RAS2 space. The computational effort is therefore substantially decreased, and the approach can be employed to study systems with large $\pi$-valence spaces. Also, the calculations are simpler because they permit the use of a unique space for the different symmetries. This approach, however, does not solve directly the valence−Rydberg mixing problems already found in CASSCF/CASPT2, leading to too high excitation energies and heavily mixed wave functions with too large orbital extensions for some valence states. As previously shown, when only the $\pi\pi^*$ MOs are included in the RAS2 space, the multistate (MS) procedure, MS-RASPT2, is required to solve the mixing and provide orthogonal states with clear valence or Rydberg mixings. In the ethene case, we have also shown that the inclusion of the $\sigma\sigma^*$ MOs in the RAS1 and RAS3 spaces (not possible in general for CAS calculations) opens new possibilities but also brings some problems. Since one cannot include in RAS2 both the $\pi\pi^*$ and $\sigma\sigma^*$ MOs, the RASPT2(SD) level of calculation is not sufficient to correctly describe the $\sigma\sigma^*$ excitations. Increasing the excitation level to SDT solved the problem in the ethene case, although this might not be a general rule. When including both $\pi$ and $\sigma$ correlation within the CI reference space, the valence−Rydberg mixing was solved at the RASPT2(SDT) level, without using the MS-RASPT2(SDT) procedure. This shows the importance of electronic correlation in defining the wave function when dealing with the valence−Rydberg mixing problem.

Calculations on different heteroaromatic organic molecules, including furan, pyrrole, and some azabenzenes and nucleobases, have shown that the most computationally advantageous RASPT2 strategy, consisting of leaving the RAS2 space empty and placing the occupied and unoccupied MOs in RAS1 and RAS3, respectively, is, in general, not particularly accurate. For low-lying states, the lack of balance between the ground and excited states caused by the improper definition of RAS2 can be partially compensated if different levels of excitations are used when defining the configurational space, in particular if using SD to compute the ground and SDT to obtain the excited state. The strategy yields poorer results for higher-lying states, mainly because of the effect of the absent Rydberg MOs and states, which should be included in the calculations to obtain accurate results.

Regarding the calculation of the first-row transition metal systems, our main goal was to analyze the effect on the excitation energies of moving the 4d correlating shell from RAS2 to RAS3. The electronic spectra of the nickel atom and the copper tetrachloride dianion have been analyzed. The main conclusion is that, overall, the RASPT2 calculations quite well reproduce the corresponding CASPT2 results (to within 0.1−0.2 eV), although a few exceptional cases were also observed, e.g., the $^1$S ($3d^{10}$) state of the nickel atom.

**Acknowledgment.** Research supported by projects CTQ2007-61260, CTQ2010-14892, and CSD2007-0010 Consolider-Ingenio in Molecular Nanoscience of the Spanish MEC/FEDER and the Generalitat Valenciana, by grants from the Flemish Science Foundation (FWO) and the Concerted Research Action of the Flemish Government (GOA) and by the Director, Office of Basic Energy Sciences, U.S. Department of Energy under Contract no. USDOE/DE-SC002183 and University of Minnesota Supercomputing Institute. S.V. thanks the University of Leuven (BOF) for financial support. Discussions with Prof. Per-Åke Malmqvist are deeply acknowledged. After the submission of this paper, one of the coauthors, Luis Serrano-Andrés, passed away unexpectedly. We would like to dedicate this paper to Luis, who will be terribly missed as a friend and as a colleague.

**Supporting Information Available:** Additional details of the calculations: some of the employed geometries, details of the symmetry restrictions used, sizes of the configurational spaces and Mulliken spin population for the states of the

copper tetrachloride dianion. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References


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

(2) Andersson, K.; Roos, B. O. *Int. J. Quantum Chem.* **1993**, *45*, 591.

(3) Merchán, M.; Serrano-Andrés, L. *J. Am. Chem. Soc.* **2003**, *125*, 8108.

(4) Hrovat, D. A.; Morokuma, K.; Borden, W. T. *J. Am. Chem. Soc.* **1994**, *116*, 1072.

(5) Lindh, R.; Persson, B. J. *J. Am. Chem. Soc.* **1994**, *116*, 4963.

(6) Moriarty, N. W.; Lindh, R.; Karlstrom, G. *Chem. Phys. Lett.* **1998**, *289*, 442.

(7) Serrano-Andrés, L.; Merchán, M.; Nebotgil, I.; Lindh, R.; Roos, B. O. *J. Chem. Phys.* **1993**, *98*, 3151.

(8) Serrano-Andrés, L.; Lindh, R.; Roos, B. O.; Merchán, M. *J. Phys. Chem.* **1993**, *97*, 9360.

(9) Serrano-Andrés, L.; Roos, B. O. *J. Am. Chem. Soc.* **1996**, *118*, 185.

(10) Fülscher, M. P.; Serrano-Andrés, L.; Roos, B. O. *J. Am. Chem. Soc.* **1997**, *119*, 6168.

(11) La Macchia, G.; Li Manni, G.; Todorova, T. K.; Brynda, M.; Aquilante, F.; Roos, B. O.; Gagliardi, L. *Inorg. Chem.* **2010**, *49*, 5216.

(12) Radoń, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824.

(13) Creve, S.; Pierloot, K.; Nguyen, M. T.; Vanquickenborne, L. G. *Eur. J. Inorg. Chem.* **1999**, 107.

(14) Persson, B. J.; Roos, B. O.; Pierloot, K. *J. Chem. Phys.* **1994**, *101*, 6810.

(15) Pierloot, K.; Persson, B. J.; Roos, B. O. *J. Phys. Chem.* **1995**, *99*, 3465.

(16) Roos, B. O.; Borin, A. C.; Gagliardi, L. *Angew. Chem., Int. Ed.* **2007**, *46*, 1469.

(17) Gagliardi, L.; Roos, B. O. *Inorg. Chem.* **2003**, *42*, 1599.

(18) Pierloot, K.; Van Praet, E.; Vanquickenborne, L. G.; Roos, B. O. *J. Phys. Chem.* **1993**, *97*, 12220.

(19) Pierloot, K.; Tsokos, E.; Vanquickenborne, L. G. *J. Phys. Chem.* **1996**, *100*, 16545.

(20) Pierloot, K.; De Kerpel, J. O. A.; Ryde, U.; Roos, B. O. *J. Am. Chem. Soc.* **1997**, *119*, 218.

(21) Pierloot, K.; De Kerpel, J. O. A.; Ryde, U.; Olsson, M.; Roos, B. O. *J. Am. Chem. Soc.* **1998**, *120*, 13156.

(22) Delabie, A.; Pierloot, K.; Groothaert, M. H.; Schoonheydt, R. A.; Vanquickenborne, L. G. *Eur. J. Inorg. Chem.* **2002**, *3*, 515.

(23) Gagliardi, L.; Roos, B. O. *Chem. Soc. Rev.* **2007**, *36*, 893.

(24) Roos, B. O.; Malmqvist, P. Å.; Gagliardi, L. *J. Am. Chem. Soc.* **2006**, *128*, 17000.

(25) Gagliardi, L. *Theor. Chem. Acc.* **2006**, *116*, 307.

(26) Pierloot, K.; van Besien, E. *J. Chem. Phys.* **2005**, *123*, 204309.

(27) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P. A.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. In *Advances in Chemical Physics, Vol XCIII*; John Wiley & Sons Inc: New York, 1996; Vol. 93, p 219.

(28) Serrano-Andrés, L.; Merchán, M.; Rubio, M.; Roos, B. O. *Chem. Phys. Lett.* **1998**, *295*, 195.

(29) Schreiber, M.; Silva, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(30) Ghigo, G.; Roos, B. O.; Malmqvist, P. A. *Chem. Phys. Lett.* **2004**, *396*, 142.

(31) Serrano-Andrés, L.; Merchán, M. *THEOCHEM* **2005**, *729*, 99.

(32) Forsberg, N.; Malmqvist, P. A. *Chem. Phys. Lett.* **1997**, *274*, 196.

(33) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M.; Molina, V. *THEOCHEM* **1996**, *388*, 257.

(34) Serrano-Andrés, L.; Merchán, M.; Lindh, R. *J. Chem. Phys.* **2005**, *122*, 104107.

(35) Finley, J.; Malmqvist, P. A.; Roos, B. O.; Serrano-Andres, L. *Chem. Phys. Lett.* **1998**, *288*, 299.

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

(37) Aquilante, F.; Pedersen, T. B.; Lindh, R.; Roos, B. O.; De Meras, A. S.; Koch, H. *J. Chem. Phys.* **2008**, *129*, 8.

(38) Aquilante, F.; Malmqvist, P. A.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694.

(39) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Pedersen, T.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.

(40) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2009**, *130*, 154107.

(41) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104.

(42) Aquilante, F.; Todorova, T. K.; Gagliardi, L.; Pedersen, T. B.; Roos, B. *J. Chem. Phys.* **2009**, *131*, 7.

(43) Malmqvist, P. Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.

(44) Roos, B. O. In *Ab Initio Methods in Quantum Chemistry*, Part II; Lawley, K. P., Ed.; Wiley: Chicester, U. K., 1987.

(45) Olsen, J.; Roos, B. O.; Jorgensen, P.; Jensen, H. J. A. *J. Chem. Phys.* **1988**, *89*, 2185.

(46) Malmqvist, P.-Å.; Rendell, A.; Roos, B. O. *J. Phys. Chem.* **1990**, *94*, 5477.

(47) Huber, S. M.; Moughal Shahi, A. R.; Aquilante, F.; Cramer, C. J.; Gagliardi, L. *J. Chem. Theory Comput.* **2009**, *5*, 2967.

(48) Moughal Shahi, A. R.; Cramer, C. J.; Gagliardi, L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10964.

(49) Widmark, P. O.; Malmqvist, P. A.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291.

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

(51) Aquilante, F.; Malmqvist, P.-A.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694.

(52) Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 11.

(53) Aquilante, F.; Pedersen, T. B.; Sanchez de Meras, A.; Koch, H. *J. Chem. Phys.* **2006**, *125*, 174101.

(54) Roos, B. O.; Lindh, R.; Malmqvist, P. A.; Veryazov, V.; Widmark, P. O. *J. Phys. Chem. A* **2005**, *109*, 6575.

(55) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742.

(56) Douglas, N.; Kroll, N. M. *Annu. Phys.* **1974**, *82*, 89.

(57) Vancoille, S.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 4011.

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

(59) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(60) Merchán, M.; Ortí, E.; Roos, B. O. *Chem. Phys. Lett.* **1994**, *226*, 27.

(61) Gwaltney, S. R.; Bartlett, R. J. *J. Chem. Phys.* **1998**, *108*, 6790.

(62) Serrano-Pérez, J. J.; Serrano-Andrés, L.; Merchán, M. *J. Chem. Phys.* **2006**, *124*, 124502.

(63) Serrano-Andrés, L.; Merchán, M.; Nebotgil, I.; Roos, B. O.; Fulscher, M. *J. Am. Chem. Soc.* **1993**, *115*, 6184.

(64) Serrano-Andrés, L.; Sánchez-Marín, J.; Nebot-Gil, I. *J. Chem. Phys.* **1992**, *97*, 7499.

(65) Christiansen, O.; Koch, H.; Halkier, A.; Jorgensen, P.; Helgaker, T.; Sánchez de Merás, A. *J. Chem. Phys.* **1996**, *105*, 6921.

(66) Halda, K.; Hättig, C.; Jorgensen, P. *J. Chem. Phys.* **2000**, *113*, 7765.

(67) Rubio, M.; Merchán, M.; Ortí, E.; Roos, B. O. *Chem. Phys.* **1994**, *179*, 395.

(68) George, G. A.; Morris, G. C. *J. Mol. Spetrosc.* **1968**, *26*, 67.

(69) Huebner, R. H.; Mielczarek, S. R.; Kuyait, C. E. *Chem. Phys. Lett.* **1972**, *16*, 464.

(70) Mikami, N.; Ito, M. *Chem. Phys. Lett.* **1975**, *31*, 472.

(71) Dick, B.; Hohlneicher, G. *Chem. Phys. Lett.* **1981**, *84*, 471.

(72) Klevens, H. B.; Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 470.

(73) Bree, A.; Trirunamachandran, T. *Mol. Phys.* **1962**, *5*, 397.

(74) Hunziker, H. E. *J. Chem. Phys.* **1972**, *56*, 400.

(75) Hunziker, H. E. *Chem. Phys. Lett.* **1969**, *3*, 504.

(76) Meyer, Y. H.; Astier, R.; Leclerq, J. M. *J. Chem. Phys.* **1972**, *56*, 801.

(77) Serrano-Andrés; L.; Roos, B. O. *Chem. Eur. J.* **1997**, *3*, 717.

(78) Flicker, W. M.; Mosher, O. A.; Kuppermann, A. *J. Chem. Phys.* **1976**, *64*, 1315.

(79) Bavia, M.; Bertinelli, F.; Taliani, C.; Zauli, C. *Mol. Phys.* **1976**, *31*, 479.

(80) Van Veen, E. H. *Chem. Phys. Lett.* **1976**, *41*, 535.

(81) Lorentzon, J.; Fülscher, M. P.; Roos, B. O. *Theor. Chim. Acta* **1995**, *92*, 67.

(82) Bolovinos, A.; Tsekeris, P.; Philis, J.; Pantos, E.; Andritsopoulus, G. *J. Mol. Spetrosc.* **1984**, *103*, 240.

(83) Fülscher, M. P.; Roos, B. O. *Theor. Chim. Acta* **1994**, *87*, 403.

(84) Innes, K. K.; Ross, I. G.; Moomaw, W. R. *J. Mol. Spetrosc.* **1988**, *132*, 49.

(85) Bolovinos, A.; Tsekeris, P.; Philis, J.; Pantos, E.; Andritsopoulus, G. *Chem. Phys.* **1990**, *147*, 19.

(86) Okuzawa, Y.; Fujii, M.; Ito, M. *Chem. Phys. Lett.* **1990**, *171*, 341.

(87) Fülscher, M. P.; Andersson, K.; Roos, B. O. *J. Phys. Chem.* **1992**, *96*, 9204.

(88) Malmqvist, P.-Å.; Roos, B. O.; Fülscher, M. P.; Rendell, A. *Chem. Phys.* **1992**, *162*, 359.

(89) Lorentzon, J.; Fülscher, M. P.; Roos, B. O. *J. Am. Chem. Soc.* **1995**, *117*, 9265.

(90) Serrano-Pérez, J. J.; González-Luque, R.; Merchán, M.; Serrano-Andrés, L. *J. Phys. Chem. B* **2007**, *111*, 11880.

(91) Fülscher, M. P.; Roos, B. O. *J. Am. Chem. Soc.* **1995**, *117*, 2089.

(92) Pierloot, K. In *Computational Organometallic Chemistry*; Cundari, T., Ed.; Marcel Dekker, Inc.: New York, 2001; p 123.

(93) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083.

(94) Linstrom, P. J.; Mallard, W. G. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology: Gaithersburg, MD, 2010.

(95) Vancoillie, S.; Chalupský, J.; Ryde, U.; Solomon, E. I.; Pierloot, K.; Neese, F.; Rulíšek, L. *J. Phys. Chem. B* **2010**, *114*, 7692.

(96) Vancoillie, S.; Malmqvist, P.-Å.; Pierloot, K. *Chem. Phys. Chem.* **2007**, *8*, 1803.

CT100478D

JCTC Journal of Chemical Theory and Computation

# DFT Calculations of Isotropic Hyperfine Coupling Constants of Nitrogen Aromatic Radicals: The Challenge of Nitroxide Radicals

L. Hermosilla, J. M. García de la Vega, C. Sieiro, and P. Calle*

*Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

Received October 26, 2010

**Abstract:** The performance of DFT methodology to predict with accuracy the isotropic hyperfine coupling constants (hfccs) of aromatic radicals containing $^{14}$N nucleus is investigated by an extensive study in which 165 hfccs, belonging to 38 radical species, are obtained from calculations with B3LYP and PBE0 functionals combined with 6-31G*, N07D, TZVP, and EPR-III basis sets, and are compared to the reported experimental data. The results indicate that the selection of the basis set is of fundamental importance in the calculation of $^{14}$N hfccs, whereas there is not so great an influence on the accurate computation of that parameter for $^{1}$H nuclei. The values of the calculated $^{14}$N coupling constants of aromatic nitroxide radicals using DFT methodology are noticeably lower than the experimental ones. A very simple relation to predict these hfccs with high accuracy is proposed on the basis of the present results, as an interesting alternative to the highly computationally demanding integrated approaches so far used.

## I. Introduction

Free radicals containing nitrogen nucleus are present in many processes of chemical, physical, and biological interest. Thus, the knowledge of their electronic distribution is very valuable to get an insight into those processes. Especially remarkable is the interest in nitroxide radical chemistry, due to their role as spin labels and spin probes, thanks to their characteristic long lives and spectromagnetic properties.[1] The nuclear hyperfine interaction, experimentally measured by electron paramagnetic resonance (EPR) spectroscopy, provides information about the electronic distribution, so the correct interpretation of the EPR spectra is of fundamental importance.[2,3] In this regard, quantum mechanical (QM) calculations can act as an effective tool for that challenge.[4] The interaction between magnetic nuclei and unpaired electrons is represented by the hyperfine tensor, which can be factored into both an isotropic (spherically symmetric) and an anisotropic (dipolar) term. The isotropic term ($a_{iso}$), so-called isotropic hyperfine coupling constant (hfcc), depends on the spin density at the nucleus position, making this property very sensitive to the level of the calculation, specifically to the electron correlation, the one-electron basis set, and the use of an adequate molecular geometry.

In previous papers,[5−8] we investigated the reliability of density functional theory (DFT) methodology to compute hfccs of different nuclei of a large number of both organic and inorganic radicals on their ground state. The main conclusion was that the best overall results are obtained when B3LYP[9,10] functional is combined with TZVP[11] or EPR-III[12,13] basis sets, yielding highly accurate values of hfccs of nuclei belonging to the three first rows. An exception was found for the $^{14}$N nuclei in which the smaller and less computationally demanding 6-31G*[14,15] basis set yields hfcc values closer to the experimental ones, probably due to the fact that it has six *d* functions instead of the five *d* functions of the TZVP and EPR-III basis sets, providing an additional *s* function to complete the *s* space.

Afterward, Barone and co-workers developed a new polarized split-valence basis set for the calculation of hfccs of second- and third-row atoms, the so-called N07D,[16,17] by adding a reduced number of polarization and diffuse functions to the 6-31G set. In order to get accurate values of the hfccs and retain, or even improve, the good performance of the parent 6-31G* basis set for other properties dominated

* Corresponding author e-mail: paloma.calle@uam.es.

10.1021/ct1006136 
Published on Web 12/02/2010

by valence orbitals, the new set was tailored by optimizing the core−valence *s* functions and reoptimizing polarization and diffuse *p* functions. Such a parametrization was made specifically for both the B3LYP and the parameter free PBE0[18] functionals. These authors reported calculations on a large set of radicals, in which the results obtained with B3LYP/N07D and PBE0/N07D combinations were compared to those produced by 6-31G*, TZVP and EPR-III basis sets, all in conjunction with B3LYP functional. The general conclusion was that both computational models, B3LYP/N07D and PBE0/N07D, provide good agreement with experimental data and predictive power at lower computational cost. For radicals containing oxygen and nitrogen atoms, they obtained poorer correlation between computations and experiments.[16] These lower correlation coefficients for $^{14}$N and $^{17}$O could be related to the reduced range of experimental data (about 30 G).

As far as we know, all previous works performed on large sets of radicals with the aim of establishing a computational protocol able to predict $a_{iso}(^{14}N)$ with high reliability, are mainly performed on nonaromatic nitrogen radicals, even though the aromatic counterparts have a lot of applications in different fields, specially nitroxide radicals, that are probably the most widely used spin probes and spin labels. The lack of a systematic theoretical study on hfccs of aromatic nitrogen radicals, which is expected to arouse a lot of interest, prompted us to investigate the performance of DFT methodology to predict with precision such constants in this work. The main goal is undertaken by performing an extensive study on a set of conjugated radicals containing $^{14}$N nucleus, in which calculation of the hfccs is carried out with the levels of theory formerly proved as the most adequate for the evaluation of this constant, that is, by employing B3LYP and PBE0 functionals combined with 6-31G*, N07D, TZVP and EPR-III basis sets on the previously optimized structures. Computed values are compared to the available experimental data by a statistical analysis. Besides the conclusions regarding the accuracy of the different methodologies on the prediction of this magnetic property, this work seeks to be a useful tool for EPR spectroscopists, since it facilitates the correct assignment of the experimental hfccs from reliable theoretical values.

## II. Computational Details

A set of 38 neutral aromatic radicals containing at least one $^{14}$N nucleus (nuclear spin $I = 1$) is considered in this study. Their schematic structures are depicted in Figure 1. The molecular geometries of the radicals on their electronic ground state are fully optimized at the B3LYP level employing the 6-31G* basis set due to its low computational cost and good results according to previous works.[5−8] Harmonic vibrational frequencies are computed at the same level of theory as the geometry optimization to confirm the nature of the stationary points. The hfccs of the radicals are evaluated on the optimized structures at five different levels of theory: PBE0/N07D, B3LYP/6-31G*, B3LYP/N07D, B3LYP/TZVP, and B3LYP/EPR-III.

The 6-31G* basis set is a small double-ζ basis plus polarization, whereas the TZVP is a DFT-optimized valence triple-ζ basis. EPR-III is a larger basis set (triple-ζ basis including diffuse functions, double *d*-polarizations, and a single set of *f*-polarization functions) optimized for the computation of hfccs by DFT methods. As previously mentioned, N07D is a polarized split-valence basis set also developed for the calculation of hfccs with an optimum compromise between reliability and computer time.[16,17] It is important to note that, although it has the same name, N07D basis set is slightly different for application with either the B3LYP or the PBE0 functional, since it has been parametrized specifically for each of them. The standard programs for the calculation of molecular structures use five *d* Cartesian Gaussian functions for TZVP and EPR-III basis sets, and six *d* functions for 6-31G* basis set. The redundant set of six *d* functions has to be employed also for N07D basis set, since it has been developed with this feature.[16,17] All calculations are carried out with the Gaussian03 software package.[19]

## III. Results and Discussion

A total of 165 hfccs for the radicals drawn in Figure 1 has been calculated, on the corresponding optimized geometries, those with available experimental data, from which 47 are $a_{iso}$ of $^{14}$N and 114 are $a_{iso}$ of $^{1}$H. The four remaining coupling constants, which correspond to $^{33}$S, $^{13}$C, and $^{17}$O nuclei, have not been taken into account in the data analysis, being the target of the paper comparison between calculated and experimental hfccs of $^{14}$N and $^{1}$H nuclei.

The name, the symmetry of the electronic ground state, and the total energies corresponding to the minimum for each radical, computed at the five different levels of theory, are shown in Table S1 in the Supporting Information (SI).

The calculated and the experimental $a_{iso}$ are summarized in Table 1. The first column corresponds to the number of each radical. The second column indicates the nuclei with their mass numbers, preceded by a number to indicate the equivalent atoms, and with a subscript to identify the nonequivalent atoms unequivocally, when necessary. The following five columns report the theoretical hfcc values obtained at the different levels of theory. In the last two columns, the experimental hfccs and their bibliographic references are given. As is well-known, the sign of $a_{iso}$ is not determined by EPR experiments; it is assigned on the basis of theoretical results. Thus, the experimental data are given as absolute values and the sign has been included just in the theoretical data. The assignment of the specified pairs of experimental hfccs of radicals **8**, **13**, and **34** has been exchanged according to the theoretical results obtained in this work.

A general inspection of the data shown in Table 1 indicates that the $a_{iso}$ of $^{1}$H are predicted in very good agreement with the experimental values, regardless of the level of theory employed. However, different conclusions are extracted from the observation of the theoretical values of $a_{iso}(^{14}N)$. In general, 6-31G* and N07D basis set present the best predictive behavior, yielding values closer to the experimental data than those obtained by EPR-III or TZVP basis sets, which tend to underestimate $a_{iso}(^{14}N)$, specially TZVP

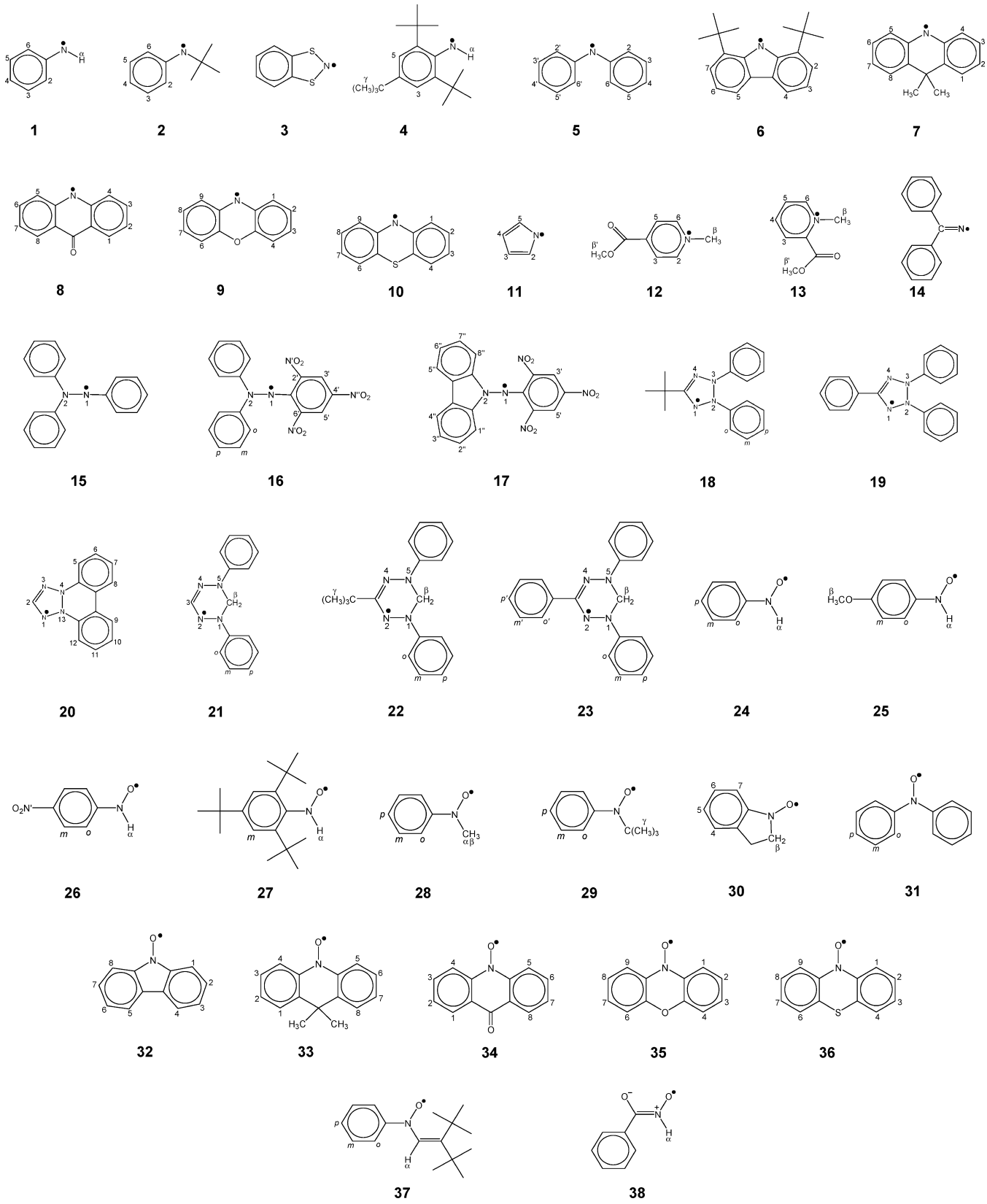



***Figure 1.*** Geometrical structures of the studied radicals.

basis set. These results are consistent with that obtained in our previous work on nitrogen coupling constants of non-aromatic radicals,[8] which pointed out that the 6-31G* basis set leads to more accurate results than TZVP and EPR-III basis sets, in spite of being smaller. The additional *s* function implicitly added when using a 6 *d* set plays a non negligible role in completing the *s* space, and thus in obtaining more accurate hfccs, in case of small or medium size basis sets (e.g., 6-31G*, N07D). However, this is not the case for larger basis sets like TZVP and EPR-III, since the results are very similar using either 5 or 6 *d* functions.[8,16,17] Accordingly, a set of 5 *d* functions is the standard for TZVP and EPR-III

***Table 1.*** Theoretical Isotropic Hyperfine Coupling Constants ($G$) of the Studied Radicals Calculated at Different Levels of Theory[a]

| no. | nuclei | $a_{iso}$ (theoretical) | | | | | experimental | |
|---|---|---|---|---|---|---|---|---|
| | | PBE0/N07D | B3LYP/6-31G* | B3LYP/N07D | B3LYP/TZVP | B3LYP/EPR-III | $a_{iso}$ | ref |
| **1** | $^{14}N$ | +9.6 | +9.3 | +9.2 | +6.1 | +7.3 | 7.95 | 20 |
| | $^{1}H_{\alpha}$ | −14.0 | −14.7 | −14.5 | −14.1 | −13.8 | 12.94 | |
| | $^{1}H_{2,6}$ | −6.7 | −6.8 | −6.7 | −6.2 | −6.3 | 6.18 | |
| | $^{1}H_{3,5}$ | +3.2 | +2.9 | +2.7 | +2.6 | +2.7 | 2.01 | |
| | $^{1}H_{4}$ | −7.7 | −7.9 | −7.8 | −7.3 | −7.6 | 8.22 | |
| **2** | $^{14}N$ | +11.3 | +10.7 | +10.7 | +7.5 | +8.6 | 9.70 | 21 |
| | $^{1}H_{2,6}$ | −6.3 | −6.4 | −6.2 | −5.7 | −5.9 | 5.84 | |
| | $^{1}H_{3,5}$ | +3.0 | +2.7 | +2.4 | +2.3 | +2.4 | 1.99 | |
| | $^{1}H_{4}$ | −7.1 | −7.3 | −7.1 | −6.6 | −6.9 | 7.09 | |
| **3** | $^{14}N$ | +12.4 | +11.3 | +11.8 | +8.7 | [b] | 11.0 | 22 |
| | $2^{33}S$ | +2.6 | +2.7 | +3.0 | +3.1 | [b] | 3.9 | |
| | $4^{1}H$ | −0.5 | −0.6 | −0.6 | −0.6 | [b] | 0.6 | |
| **4** | $^{14}N$ | −8.4 | +8.6 | +8.1 | +5.5 | +6.4 | 6.70 | 23 |
| | $^{1}H_{\alpha}$ | −12.4 | −13.0 | −12.8 | −12.4 | −12.2 | 11.75 | |
| | $^{1}H_{3,5}$ | +3.1 | +2.7 | +2.7 | +2.4 | +2.5 | 1.89 | |
| | $9^{1}H_{\gamma}$ | +0.2 | +0.3 | +0.3 | +0.3 | +0.3 | 0.27 | |
| **5** | $^{14}N$ | +9.9 | +9.3 | +9.2 | +6.5 | +7.4 | 8.80 | 24 |
| | $^{1}H_{2,2',6,6'}$ | −4.3 | −4.4 | −4.2 | −3.9 | −4.0 | 3.68 | |
| | $^{1}H_{3,3',5,5'}$ | +2.3 | +2.1 | +2.0 | +1.9 | +2.0 | 1.52 | |
| | $^{1}H_{4,4'}$ | −4.8 | −4.8 | −4.8 | −4.4 | −4.6 | 4.28 | |
| **6** | $^{14}N$ | +8.6 | +8.1 | +8.0 | +5.5 | +6.3 | 6.97 | 25 |
| | $^{1}H_{2,7}$ | +1.6 | +1.4 | +1.3 | +1.2 | +1.3 | 0.89 | |
| | $^{1}H_{3,6}$ | −4.4 | −4.4 | −4.4 | −4.0 | −4.2 | 4.30 | |
| | $^{1}H_{4,5}$ | +1.2 | +0.4 | +0.4 | +0.4 | +0.1 | 0.14 | |
| **7** | $^{14}N$ | +9.2 | +8.4 | +8.6 | +5.9 | +6.8 | 8.00 | 24 |
| | $^{1}H_{1,3,6,8}$ | +2.2 | +2.0 | +1.9 | +1.7 | +1.8 | 1.28 | |
| | $^{1}H_{2,7}$ | −5.0 | −5.0 | −4.9 | −4.6 | −4.8 | 4.52 | |
| | $^{1}H_{4,5}$ | −4.4 | −4.4 | −4.3 | −3.9 | −4.1 | 3.67 | |
| **8** | $^{14}N$ | +8.0 | +7.3 | +7.4 | +5.1 | +5.9 | 6.98 | 24 |
| | $^{1}H_{1,8}$ | +1.6 | +1.3 | +1.2 | +1.1 | +1.1 | 0.76[c] | |
| | $^{1}H_{2,7}$ | −4.5 | −4.5 | −4.4 | −4.1 | −4.3 | 4.12 | |
| | $^{1}H_{3,6}$ | +2.2 | +1.9 | +1.8 | +1.6 | +1.7 | 1.27[c] | |
| | $^{1}H_{4,5}$ | −4.4 | −4.3 | −4.2 | −3.9 | −4.0 | 3.67 | |
| **9** | $^{14}N$ | +8.8 | +8.1 | +8.2 | +5.7 | +6.6 | 8.03 | 26 |
| | $^{1}H_{1,9}$ | −3.7 | −3.7 | −3.6 | −3.3 | −3.4 | 2.88 | |
| | $^{1}H_{2,8}$ | +1.5 | +1.3 | +1.2 | +1.1 | +1.1 | 0.97 | |
| | $^{1}H_{3,7}$ | −4.4 | −4.4 | −4.3 | −4.0 | −4.2 | 3.97 | |
| | $^{1}H_{4,6}$ | +1.6 | +1.4 | +1.3 | +1.2 | +1.3 | 0.65 | |
| **10** | $^{14}N$ | +8.1 | +7.5 | +7.5 | +5.3 | [b] | 7.05 | 27 |
| | $^{1}H_{1,9}$ | −3.7 | −3.7 | −3.5 | −3.2 | [b] | 2.85 | |
| | $^{1}H_{2,4,6,8}$ | +1.8 | +1.5 | +1.4 | +1.3 | [b] | 0.95 | |
| | $^{1}H_{3,7}$ | −4.2 | −4.3 | −4.2 | −3.9 | [b] | 3.66 | |
| **11** | $^{14}N$ | −3.2 | −2.4 | −2.6 | −2.3 | −2.4 | 2.91 | 28 |
| | $^{1}H_{2,5}$ | −13.5 | −14.0 | −13.8 | −12.9 | −13.4 | 13.26 | |
| | $^{1}H_{3,4}$ | −3.2 | −3.4 | −3.4 | −3.2 | −3.3 | 3.55 | |
| **12** | $^{14}N$ | +6.3 | +5.8 | +5.9 | +4.5 | +5.0 | 6.25 | 29 |
| | $^{1}H_{2,6}$ | −4.7 | −5.1 | −4.8 | −4.5 | −4.6 | 3.55 | |
| | $^{1}H_{3,5}$ | +0.4 | +0.2 | +0.1 | +0.1 | +0.1 | 0.8 | |
| | $3^{1}H_{\beta}$ | +5.2 | +5.8 | +5.1 | +5.4 | +5.7 | 5.55 | |
| | $3^{1}H_{\beta'}$ | +0.9 | +0.9 | +0.9 | +0.9 | +1.0 | 0.8 | |
| **13** | $^{14}N$ | +6.2 | +5.6 | +5.7 | +4.3 | +4.8 | 6.58 | 30 |
| | $^{1}H_{3}$ | +1.8 | +1.6 | +1.5 | +1.4 | +1.4 | 0.94 | |
| | $^{1}H_{4}$ | −6.7 | −7.2 | −6.9 | −6.4 | −6.6 | 6.28 | |
| | $^{1}H_{5}$ | −2.5 | −2.8 | −2.8 | −2.6 | −2.7 | 2.54[c] | |
| | $^{1}H_{6}$ | −1.4 | −1.6 | −1.5 | −1.3 | −1.4 | 1.40[c] | |
| | $3^{1}H_{\beta}$ | +5.8 | +6.3 | +6.0 | +5.9 | +6.4 | 5.64 | |
| | $3^{1}H_{\beta'}$ | +1.0 | +0.5 | +1.1 | +1.1 | +1.2 | 0.94 | |
| **14** | $^{14}N$ | +11.5 | +11.0 | +11.8 | +6.7 | +8.6 | 10 | 31 |
| | $4^{1}H_{o},4^{1}H_{m}$ | +0.3 | +0.4 | +0.4 | +0.4 | +0.4 | 0.37 | |
| | $2^{1}H_{p}$ | +0.2 | +0.2 | +0.2 | +0.2 | +0.2 | <0.2 | |
| **15** | $^{14}N_{1}$ | +11.0 | +10.7 | +10.4 | +7.8 | +8.7 | 9.05 | 32 |
| | $^{14}N_{2}$ | +4.1 | +4.1 | +3.9 | +2.6 | +3.1 | 4.28 | |
| **16** | $^{14}N_{1}$ | +12.4 | +12.1 | +12.2 | +10.5 | +11.1 | 9.74 | 33 |
| | $^{14}N_{2}$ | +7.4 | +7.0 | +6.9 | +5.1 | +5.7 | 7.95 | |
| | $2^{14}N'$ | −0.4 | −0.3 | −0.3 | −0.4 | −0.3 | 0.39 | 34 |
| | $^{14}N''$ | −0.4 | −0.3 | −0.3 | −0.4 | −0.3 | 0.48 | |
| | $^{1}H_{3',5'}$ | +1.4 | +1.2 | +1.2 | +1.2 | +1.2 | 1.06 | 35 |
| | $4^{1}H_{o}$ | −1.8 | −1.7 | −1.7 | −1.6 | −1.6 | 1.55 | |
| | $4^{1}H_{m}$ | +1.0 | +0.9 | +0.9 | +0.8 | +0.9 | 0.73 | |
| | $2^{1}H_{p}$ | −1.7 | −1.7 | −1.7 | −1.6 | −1.7 | 1.58 | |
| **17** | $^{14}N_{1}$ | +14.0 | +13.4 | +13.6 | +11.9 | +12.4 | 11.1 | 36 |
| | $^{14}N_{2}$ | +5.3 | +5.0 | +5.0 | +3.5 | +4.1 | 6.0 | |
| | $^{1}H_{3',5'}$ | +1.5 | +1.4 | +1.4 | +1.3 | +1.4 | 1.17 | |
| | $^{1}H_{1'',8''}$ | −2.2 | −2.2 | −2.1 | −2.0 | −2.1 | 1.92 | |
| | $^{1}H_{2'',7''}$ | +0.8 | +0.7 | +0.7 | +0.5 | +0.7 | 0.53 | |
| | $^{1}H_{3'',6''}$ | −2.0 | −2.1 | −2.0 | −1.9 | −2.0 | 1.81 | |
| | $^{1}H_{4'',5''}$ | +0.7 | +0.6 | +0.5 | +0.7 | +0.5 | 0.41 | |

***Table 1.*** Continued

| no. | nuclei | $a_{iso}$ (theoretical) PBE0/N07D | B3LYP/6-31G* | B3LYP/N07D | B3LYP/TZVP | B3LYP/EPR-III | experimental $a_{iso}$ | ref |
|---|---|---|---|---|---|---|---|---|
| **18** | $^{14}N_{1,4}$ | +6.5 | +6.2 | +6.1 | +4.6 | +5.1 | 5.7 | 37 |
| | $^{14}N_{2,3}$ | +7.1 | +6.8 | +6.8 | +5.5 | +5.9 | 7.5 | |
| | $4^1H_o$,$2^1H_p$ | −1.1 | −1.1 | −1.1 | −1.0 | −1.1 | 0.95 | |
| | $4^1H_m$ | +0.7 | +0.6 | +0.6 | +0.6 | +0.6 | 0.5 | |
| **19** | $^{14}N_{1,4}$ | +6.5 | +6.2 | +6.1 | +4.5 | +5.1 | 5.6 | 38 |
| | $^{14}N_{2,3}$ | +7.1 | +6.9 | +7.0 | +5.7 | +6.1 | 7.5 | |
| **20** | $^{14}N_{1,3}$ | +4.2 | +4.1 | +3.9 | +2.8 | +3.2 | 3.85 | 39 |
| | $^{14}N_{4,13}$ | +5.7 | +5.1 | +5.2 | +3.8 | +4.4 | 7.7 | |
| | $^1H_{5,7,10,12}$ | −2.6 | −2.7 | −2.6 | −2.4 | −2.5 | 1.9 | |
| | $^1H_{6,11}$ | +0.9 | +0.8 | +0.7 | +0.6 | +0.7 | 0.4 | |
| | $^1H_{8,9}$ | +0.5 | +0.4 | +0.3 | +0.3 | +0.3 | 0.4 | |
| **21** | $^{14}N_{1,2,4,5}$ | +5.8 | +5.7 | +5.5 | +4.1 | +4.6 | 6.0 | 40 |
| | $^1H_3$ | +2.3 | +2.0 | +1.9 | +1.7 | +1.8 | 0.72 | 41 |
| | $4^1H_o$ | −1.4 | −1.4 | −1.4 | −1.2 | −1.3 | 1.10 | |
| | $4^1H_m$ | +0.7 | +0.6 | +0.6 | +0.5 | +0.6 | 0.40 | |
| | $2^1H_p$ | −1.5 | −1.5 | −1.5 | −1.4 | −1.5 | 1.16 | |
| **22** | $^{14}N_{1,2,4,5}$ | +5.7 | +5.5 | +5.4 | +4.0 | +4.5 | 5.9 | 42 |
| | $4^1H_o$,$2^1H_p$ | −1.4 | −1.5 | −1.4 | −1.1 | −1.4 | 1.08 | 41 |
| | $4^1H_m$ | +0.7 | +0.6 | +0.6 | +0.5 | +0.6 | 0.40 | |
| | $2^1H_\beta$ | −0.1 | −0.2 | −0.1 | −0.1 | −0.1 | 0.08 | |
| | $9^1H_\gamma$ | +0.1 | +0.1 | +0.1 | +0.1 | +0.1 | 0.11 | |
| **23** | $^{14}N_{1,2,4,5}$ | +5.7 | +5.6 | +5.4 | +4.0 | +4.4 | 5.79 | 43 |
| | $4^1H_o$ | −1.4 | −1.5 | −1.4 | −1.3 | −1.4 | 1.12 | 44 |
| | $4^1H_m$ | +0.4 | +0.7 | +0.6 | +0.6 | +0.6 | 0.43 | |
| | $2^1H_p$ | −1.6 | −1.6 | −1.6 | −1.4 | −1.5 | 1.20 | |
| | $2^1H_{o'}$ | +0.9 | +0.8 | +0.8 | +0.7 | +0.7 | 0.43 | |
| | $2^1H_{m'}$ | −0.4 | −0.3 | −0.3 | −0.3 | −0.3 | 0.16 | |
| | $^1H_{p'}$ | +0.7 | +0.6 | +0.6 | +0.5 | +0.6 | 0.31 | |
| | $2^1H_\beta$ | +0.2 | −0.01 | +0.02 | +0.05 | +0.1 | 0.03 | |
| **24** | $^{14}N$ | +7.3 | +6.8 | +7.0 | +4.7 | +5.6 | 9.75 | 45 |
| | $^1H_\alpha$ | −11.9 | −12.7 | −13.0 | −12.3 | −12.6 | 12.75 | |
| | $2^1H_o$,$^1H_p$ | −3.3 | −3.2 | −3.3 | −3.0 | −3.2 | 3.00 | |
| | $2^1H_m$ | +1.5 | +1.4 | +1.3 | +1.2 | +1.3 | 1.00 | |
| **25** | $^{14}N$ | +7.8 | +7.2 | +7.3 | +5.0 | +5.9 | 10.15 | 45 |
| | $^1H_\alpha$ | −12.2 | −12.9 | −13.3 | −12.5 | −12.9 | 13.75 | |
| | $2^1H_o$ | −3.3 | −3.3 | −3.3 | −3.0 | −3.1 | 3.37 | |
| | $2^1H_m$ | +1.4 | +1.3 | +1.2 | +1.1 | +1.2 | 1.00 | |
| | $3^1H_\beta$ | +0.4 | +0.4 | +0.5 | +0.4 | +0.5 | 0.50 | |
| **26** | $^{14}N$ | +6.3 | +5.9 | +6.0 | +3.9 | +4.7 | 7.50 | 45 |
| | $^{14}N'$ | −0.8 | −0.6 | −0.7 | −0.7 | −0.7 | 1.85 | |
| | $^1H_\alpha$ | −11.0 | −11.7 | −12.0 | −11.2 | −11.6 | 10.10 | |
| | $2^1H_o$ | −3.1 | −3.1 | −3.0 | −2.8 | −2.9 | 3.00 | |
| | $2^1H_m$ | +1.5 | +1.3 | +1.3 | +1.2 | +1.3 | 0.85 | |
| **27** | $^{14}N$ | +9.6 | +9.1 | +9.1 | +6.6 | +7.5 | 11.65 | 46 |
| | $^1H_\alpha$ | −13.1 | −13.9 | −13.9 | −13.5 | −13.6 | 12.96 | |
| | $2^1H_m$ | +1.2 | +1.1 | +1.1 | +1.1 | +1.1 | 1.03 | |
| **28** | $^{14}N$ | +8.9 | +7.8 | +8.3 | +5.8 | +6.8 | 10.65 | 47 |
| | $2^1H_o$,$^1H_p$ | −3.1 | −3.1 | −3.1 | −2.8 | −3.0 | 2.75 | |
| | $2^1H_m$ | +1.5 | +1.3 | +1.2 | +1.1 | +1.2 | 1.01 | |
| | $3^1H_\beta$ | +7.9 | +8.4 | +8.5 | +8.3 | +9.0 | 9.69 | |
| | $^{13}C_\alpha$ | −6.0 | −5.3 | −5.8 | −6.2 | −6.0 | 6.0 | 48 |
| **29** | $^{14}N$ | +9.8 | +8.8 | +9.1 | +6.7 | +7.6 | 12.08 | 49 |
| | $2^1H_o$ | −2.9 | −2.9 | −2.9 | −2.6 | −2.7 | 2.09 | |
| | $2^1H_m$ | +1.4 | +1.2 | +1.2 | +1.1 | +1.2 | 0.89 | |
| | $^1H_p$ | −2.9 | −2.9 | −2.9 | −2.7 | −2.8 | 2.29 | |
| | $9^1H_\gamma$ | +0.03 | +0.1 | +0.1 | +0.1 | +0.1 | 0.09 | |
| **30** | $^{14}N$ | +8.9 | +7.7 | +8.3 | +5.8 | +6.8 | 11.75 | 50 |
| | $^1H_{4,6}$ | +1.6 | +1.4 | +1.3 | +1.2 | +1.3 | 1.00 | |
| | $^1H_{5,7}$ | −3.7 | −3.8 | −3.7 | −3.4 | −3.6 | 3.74 | |
| | $2^1H_\beta$ | +14.6 | +15.0 | +15.6 | +14.9 | +16.5 | 18.60 | |
| **31** | $^{14}N$ | +8.4 | +7.5 | +7.8 | +5.5 | +6.4 | 9.66 | 51 |
| | $4^1H_o$,$2^1H_p$ | −2.1 | −2.1 | −2.1 | −1.9 | −2.0 | 1.83 | |
| | $4^1H_m$ | +1.2 | +1.1 | +1.0 | +1.0 | +1.0 | 0.79 | |
| **32** | $^{14}N$ | +6.0 | +5.2 | +5.6 | +3.6 | +4.4 | 6.65 | 52 |
| | $^1H_{1,3,6,8}$ | −2.4 | −2.4 | −2.4 | −2.3 | −2.4 | 2.30 | |
| | $^1H_{2,4,5,7}$ | +0.8 | +0.8 | +0.7 | +0.7 | +0.7 | 0.55 | |
| | $^{17}O$ | −17.3 | −15.3 | −16.4 | −10.2 | −13.3 | 16.5 | |
| **33** | $^{14}N$ | +7.9 | +6.8 | +7.3 | +5.1 | +6.0 | 8.75 | 52 |
| | $^1H_{1,3,6,8}$ | +1.2 | +1.1 | +1.0 | +0.9 | +1.0 | 0.75 | |
| | $^1H_{2,4,5,7}$ | −2.6 | −2.6 | −2.6 | −2.4 | −2.5 | 2.30 | |
| | $^{17}O$ | −15.7 | −14.9 | −14.8 | −9.4 | −12.0 | 16.6 | |
| **34** | $^{14}N$ | +6.3 | +5.4 | +5.8 | +3.9 | +4.6 | 6.89 | 24 |
| | $^1H_{1,3,6,8}$ | +1.1 | +1.0 | +1.0 | +0.9 | +0.9 | 0.69 | |
| | $^1H_{2,7}$ | −2.2 | −2.2 | −2.2 | −2.0 | −2.1 | 2.03[c] | |
| | $^1H_{4,5}$ | −2.3 | −2.3 | −2.3 | −2.1 | −2.2 | 2.11[c] | |
| **35** | $^{14}N$ | +8.1 | +7.0 | +7.4 | +5.2 | +6.1 | 9.50 | 52 |
| | $^1H_{1,3,7,9}$ | −2.6 | −2.7 | −2.6 | −2.4 | −2.6 | 2.40 | |
| | $^1H_{2,4,6,8}$ | +0.9 | +0.7 | +0.7 | +0.6 | +0.6 | 0.50 | |

***Table 1.*** Continued

| no. | nuclei | $a_{iso}$ (theoretical) PBE0/N07D | B3LYP/6-31G* | B3LYP/N07D | B3LYP/TZVP | B3LYP/EPR-III | experimental $a_{iso}$ | ref |
|---|---|---|---|---|---|---|---|---|
| **36** | $^{14}$N | +8.4 | +7.3 | +7.7 | +5.6 | [b] | 9.01 | 53 |
| | $^{1}H_{1,3,7,9}$ | −2.4 | −2.5 | −2.4 | −2.2 | [b] | 2.20 | |
| | $^{1}H_{2,8}$ | +1.1 | +0.9 | +0.9 | +0.8 | [b] | 0.63 | |
| | $^{1}H_{4,6}$ | +1.0 | +0.8 | +0.8 | +0.7 | [b] | 0.50 | |
| **37** | $^{14}$N | +8.5 | +7.6 | +7.8 | +5.6 | +6.4 | 10.0 | 54 |
| | $^{1}H_{\alpha}$ | +10.3 | +10.9 | +10.8 | +10.3 | +11.2 | 13.6 | |
| | $2^{1}H_{o}$,$^{1}H_{p}$ | −2.5 | −2.5 | −2.4 | −2.2 | −2.4 | 2.5 | |
| | $2^{1}H_{m}$ | +1.2 | +1.1 | +1.0 | +0.9 | +1.0 | 0.9 | |
| **38** | $^{14}$N | +4.4 | +4.5 | +4.4 | +2.5 | +3.2 | 6.10 | 55 |
| | $^{1}H_{\alpha}$ | −9.4 | −10.4 | −10.5 | −9.9 | −10.2 | 10.40 | |

[a] All calculations have been carried out on the geometries optimized at B3LYP/6-31G* level of theory. [b] EPR-III basis set is not parametrized for the sulfur nucleus. [c] The assignment of these experimental hfccs has been exchanged taking into account the present theoretical calculations.

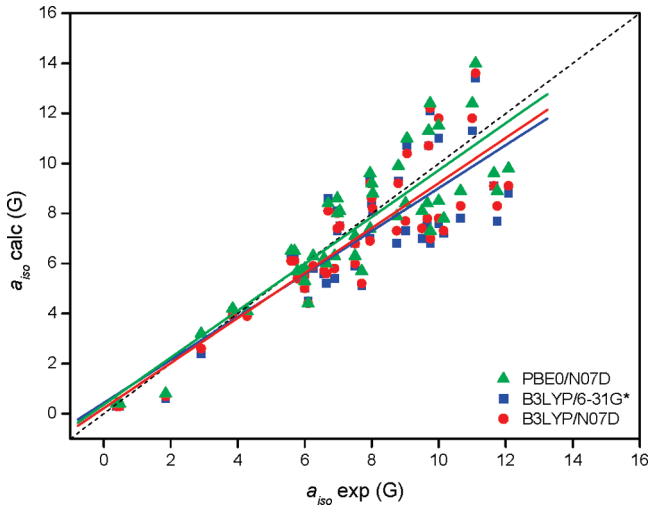



***Figure 2.*** Plot of theoretical vs experimental $a_{iso}$ for $^{14}$N nuclei of all studied radicals, calculated at B3LYP/6-31G*, B3LYP/N07D, and PBE0/N07D levels of theory. Linear fits are represented by solid lines.

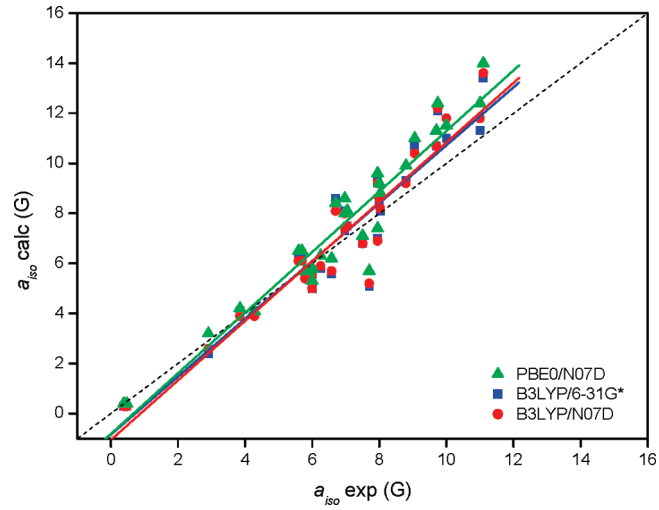



***Figure 3.*** Plot of theoretical vs experimental $a_{iso}$ for $^{14}$N nuclei of radicals **1**−**23**, calculated at B3LYP/6-31G*, B3LYP/N07D, and PBE0/N07D levels of theory. Linear fits are represented by solid lines.

basis sets, while 6 *d* functions is mandatory for 6-31G* and N07D basis sets. This result can be clearly observed for a set of ten radicals in Table S2 in the SI, which lists the differences between the $a_{iso}$($^{14}$N) calculated with 5 and 6 *d* functions. From these data, we can conclude that large basis set are not affected by the number of *d* functions (maximum differences of 0.3 G). However, for calculations with medium size basis set, the average differences are around 2G.

In order to get a deeper insight into the results obtained for the nitrogen coupling constants, we have considered in the further analysis only the hfcc data obtained with the N07D and 6-31G* basis sets, because the three combinations, PBE0/N07D, B3LYP/6-31G*, and B3LYP/N07D compute values of $^{14}$N hfccs more reliable than those obtained for combinations with EPR-III and TZVP basis sets. However, all the data obtained at the five calculation levels are displayed in Figure S1 in the SI.

Figure 2 represents the calculated vs experimental $a_{iso}$ of $^{14}$N for the 38 species, at the three calculation levels above indicated. Although this figure shows an important scattering of the points, interesting conclusions can be achieved if the set of 38 radicals is divided in two subsets, one of them corresponding to the non nitroxide-type radicals, species **1**−**23**, and the other one to the nitroxide-type radicals, species **24**−**38**.

As can be observed in Figures 3 and 4, $a_{iso}$($^{14}$N) of the non nitroxide-type radical subset are predicted quite accurately at the three calculation levels, whereas the $^{14}$N hfccs obtained for nitroxide-type radicals are noteworthy underestimated. This result had been previously observed by other authors and analyzed in the terms of the limits of the static gas-phase DFT approaches.[56] In the case of the nitroxide radicals, there are two main critical geometric parameters, namely the improper dihedral angle corresponding to the out-of-plane motion of the NO moiety, and the N−O bond length; a nearly planar environment of nitrogen leads to the lack of any contribution of nitrogen *s* orbitals to the SOMO with the consequent reduction of hfcc. The computation of the magnetic parameter along the trajectory provided by Molecular Dynamics (MD) runs may account for this effect. Indeed, several works[56−60] have shown that the computed $a_{iso}$($^{14}$N) for the planar structures predicted by static models are lower than the values obtained for the averaged ones obtained by molecular dynamic simulations, that are slightly pyramidal. As a matter of fact, all of the nitroxide radicals studied herein are predicted to have optimized geometries

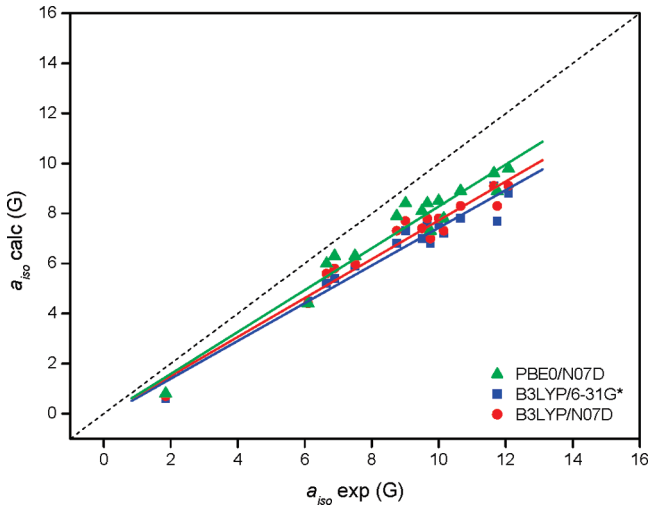



***Figure 4.*** Plot of theoretical vs experimental $a_{iso}$ for $^{14}$N nuclei of radicals **24**−**38**, calculated at B3LYP/6-31G*, B3LYP/N07D, and PBE0/N07D levels of theory. Linear fits are represented by solid lines.

with almost planar environment of the nitrogen, which could explain the eventual reduction of the hfcc.

For a better comparison of the general performance and accuracy of the different computational schemes in the prediction of hfccs, a regression analysis was carried out for the five calculation levels included in Table 1. Different parameters of the regression analysis: intercept, slope, correlation coefficient ($R^2$) of the least-squares fit, as well as the number of data ($N$), range (minimum and maximum absolute values), absolute error, maximum and mean percent error ($E$) between calculated and experimental values, mean absolute deviation (MAD), and the ratio range/MAD are collected in Table 2. The analysis was made considering $a_{iso}(^{14}N)$ and the $a_{iso}(^{1}H)$ together and separately, in all the radicals and in the two subsets, non nitroxide and nitroxide-type radicals, as indicate the sections shown in Table 2. The number of data is high enough to infer general conclusions as regards the prediction of this parameter on nitrogen aromatic radicals.

The correlation coefficient ($R^2$) for the five linear fits considering all the nuclei in the 38 radicals (section *i* in Table 2) is higher than 0.87, being the values of the calculations with B3LYP/TZVP and B3LYP/EPR-III the poorest. The $R^2$ values for the calculations with the other three levels of theory (N07D and 6-31G* basis sets) are higher (>0.93) and quite similar between each other. The same qualitative tendency is observed when non nitroxide and nitroxide-type radicals are analyzed separately (sections ii and iii in Table 2).

Analyzing the values calculated for $a_{iso}(^{14}N)$, the lowest $R^2$ value corresponds to the fits where all radical are considered together, $0.65 < R^2 < 0.77$. Correlation coefficient becomes better when both, non nitroxide and nitroxide type radical subsets, are considered separately, arising values higher than 0.9 for almost all fittings (sections v and vi in Table 2).

As expected, for $a_{iso}(^{1}H)$ the correlation coefficient of the five linear fits are very high (>0.96) without significant differences between the two types of radicals, and the best result always corresponds to the B3LYP/EPR-III combination (sections vii−ix in Table 2).

The analysis of the slopes of the least-squares fits also reveals the influence of the basis set and the type of radical, especially in the calculation of the $a_{iso}(^{14}N)$. The value of the slopes in which all nuclei ($^{1}H$ and $^{14}N$) have been taken in account are considerably far from the unit for TZVP and EPR-III basis sets, obtaining similar slope values for fittings with N07D and 6-31G* basis sets, and closer to the unit value. According to this parameter, the agreement between theoretical and experimental $a_{iso}$ is much better for non nitroxide-type radicals than for nitroxide-type radicals, since the slopes are much closer to the unity in the former. Especially remarkable is the deviation of the values predicted by the B3LYP/TZVP scheme, which slope for nitroxide-type radicals is ca. 0.70.

Analyzing the values of the $^{14}$N hfccs, it is observed that the slopes are scattered from 0.57 to 1.21 (sections v−vi in Table 2) and these values depend strongly on the kind of compounds considered, that is, all radicals together or any of the subsets, nitroxide or non-nitroxide type radicals, separately. So, the values of the slopes for the three lower-cost levels of theory are noticeably higher than the unit for radicals **1**−**23** and smaller than one for radicals **24**−**38**, resulting in slopes closer to the unity when all radicals are analyzed as a whole. However, the bad values of the correlation coefficient for these cases is a signal of the inconsistency of these data, thus, both groups must to be analyzed separately.

The value of the absolute error is not representative of the accuracy of a computational level for the prediction of hfccs since it depends on the range. Nevertheless, comparison of the maximum absolute errors within each subset is coherent. In that regard, we should stress that the values corresponding to the calculations with TZVP and EPR-III basis sets are higher than those with 6-31G* and N07D basis sets, analyzing $^{1}H$ and $^{14}N$ together or separately, in the 38 radicals or in the two subsets, non-nitroxide and nitroxide type radicals, except in the analysis of the $^{1}H$ nucleus, for which notable discrepancies do not become apparent among the five combinations, corresponding lower values of the maximum absolute errors to EPR-III calculations.

The parameter that can be considered as reference for the relative comparison of the accuracy of the different methodologies in the computation of hfccs respect to the corresponding experimental values is the ratio between the range and the mean absolute deviation (range/MAD). The most remarkable result is the lower values of that ratio for the calculations with the two larger basis sets in any subset where $^{14}$N hfccs have been included, sections i−vi in Table 2, pointing out that B3LYP/TZVP and B3LYP/EPR-III levels are not suitable for the prediction of $a_{iso}(^{14}N)$ of aromatic radicals in general. However, the opposite result is obtained for the computation of the $^{1}H$ coupling constant, as the range/MAD values for the combinations with 6-31G* and N07D basis sets are lower. In this case, the ratio for the PBE0/N07D level is the poorest, being the B3LYP/N07D combination the best one of the three less computational demanding

***Table 2.*** Regression Analysis for Predictions of Isotropic Hyperfine Coupling Constants (*G*) of the Studied Radicals

| level of theory[a] | intercept | slope | $R^2$ | *N* | $\|a_{iso}\|_{min}$ | $\|a_{iso}\|_{max}$ | max. absolute error | average *E*%[b] | max *E*% | MAD[c] | range/MAD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (i) all nuclei, all radicals | | | | | | | | | | | |
| PBE0/N07D | 0.5389 | 0.9071 | 0.9421 | 161 | 0.03 | 14.6 | 4.0 | 37 | 757 | 0.70 | 20.93 |
| B3LYP/6-31G* | 0.4400 | 0.8994 | 0.9306 | 161 | 0.01 | 15.3 | 4.1 | 27 | 186 | 0.71 | 21.14 |
| B3LYP/N07D | 0.3562 | 0.9181 | 0.9417 | 161 | 0.02 | 16.4 | 3.5 | 23 | 186 | 0.64 | 24.20 |
| B3LYP/TZVP | 0.3468 | 0.7547 | 0.8733 | 161 | 0.05 | 14.9 | 6.0 | 24 | 186 | 0.96 | 15.41 |
| B3LYP/EPR-III | 0.3398 | 0.8254 | 0.9120 | 151 | 0.1 | 16.5 | 5.0 | 23 | 233 | 0.79 | 20.82 |
| (ii) all nuclei, radicals **1**−**23** | | | | | | | | | | | |
| PBE0/N07D | 0.2702 | 1.0539 | 0.9838 | 105 | 0.1 | 14.0 | 2.9 | 43 | 757 | 0.61 | 22.94 |
| B3LYP/6-31G* | 0.1902 | 1.0429 | 0.9837 | 105 | 0.01 | 14.7 | 2.6 | 30 | 186 | 0.55 | 26.58 |
| B3LYP/N07D | 0.1279 | 1.0451 | 0.9848 | 105 | 0.02 | 14.5 | 2.5 | 25 | 186 | 0.51 | 28.50 |
| B3LYP/TZVP | 0.2311 | 0.8365 | 0.9577 | 105 | 0.05 | 14.1 | 3.9 | 24 | 186 | 0.69 | 20.40 |
| B3LYP/EPR-III | 0.2135 | 0.9099 | 0.9737 | 99 | 0.1 | 13.8 | 3.3 | 25 | 233 | 0.55 | 24.96 |
| (iii) all nuclei, radicals **24**−**38** | | | | | | | | | | | |
| PBE0/N07D | 0.5693 | 0.8029 | 0.9728 | 56 | 0.03 | 14.6 | 4.0 | 27 | 100 | 0.87 | 16.84 |
| B3LYP/6-31G* | 0.4369 | 0.7991 | 0.9426 | 56 | 0.1 | 15.0 | 4.1 | 23 | 68 | 1.00 | 14.87 |
| B3LYP/N07D | 0.3719 | 0.8284 | 0.9502 | 56 | 0.1 | 15.6 | 3.5 | 20 | 62 | 0.90 | 17.26 |
| B3LYP/TZVP | 0.2776 | 0.7010 | 0.8462 | 56 | 0.1 | 14.9 | 6.0 | 23 | 62 | 1.48 | 10.00 |
| B3LYP/EPR-III | 0.2743 | 0.7727 | 0.8913 | 52 | 0.1 | 16.5 | 5.0 | 20 | 62 | 1.24 | 13.20 |
| (iv) $^{14}$N nucleus, all radicals | | | | | | | | | | | |
| PBE0/N07D | 0.3694 | 0.9362 | 0.7700 | 47 | 0.4 | 14.0 | 2.9 | 15 | 57 | 1.15 | 11.78 |
| B3LYP/6-31G* | 0.4328 | 0.8581 | 0.7110 | 47 | 0.3 | 13.4 | 4.1 | 18 | 68 | 1.32 | 9.94 |
| B3LYP/N07D | 0.2202 | 0.8997 | 0.7500 | 47 | 0.3 | 13.6 | 3.5 | 17 | 62 | 1.24 | 10.73 |
| B3LYP/TZVP | 0.1689 | 0.6482 | 0.6580 | 47 | 0.4 | 11.9 | 6.0 | 32 | 62 | 2.54 | 4.52 |
| B3LYP/EPR-III | 0.2010 | 0.7323 | 0.7140 | 44 | 0.3 | 12.4 | 5.0 | 24 | 62 | 1.91 | 6.35 |
| (v) $^{14}$N nucleus, radicals **1**−**23** | | | | | | | | | | | |
| PBE0/N07D | −0.8090 | 1.2101 | 0.9176 | 31 | 0.4 | 14.0 | 2.9 | 13 | 27 | 0.96 | 14.19 |
| B3LYP/6-31G* | −0.8441 | 1.1566 | 0.9021 | 31 | 0.3 | 13.4 | 2.6 | 13 | 38 | 0.81 | 16.11 |
| B3LYP/N07D | −1.0486 | 1.1880 | 0.9111 | 31 | 0.3 | 13.6 | 2.5 | 13 | 38 | 0.84 | 15.91 |
| B3LYP/TZVP | −0.8054 | 0.8786 | 0.8396 | 31 | 0.4 | 11.9 | 3.9 | 25 | 51 | 1.74 | 6.63 |
| B3LYP/EPR-III | −0.9376 | 0.9943 | 0.8798 | 29 | 0.3 | 12.4 | 3.3 | 18 | 43 | 1.16 | 10.44 |
| (vi) $^{14}$N nucleus, radicals **24**−**38** | | | | | | | | | | | |
| PBE0/N07D | −0.0806 | 0.8362 | 0.9364 | 16 | 0.8 | 9.8 | 2.9 | 19 | 57 | 1.53 | 5.87 |
| B3LYP/6-31G* | −0.1105 | 0.7536 | 0.9500 | 16 | 0.6 | 9.1 | 4.1 | 28 | 68 | 2.30 | 3.70 |
| B3LYP/N07D | −0.0399 | 0.7767 | 0.9487 | 16 | 0.7 | 9.1 | 3.5 | 24 | 62 | 2.02 | 4.16 |
| B3LYP/TZVP | −0.3337 | 0.5745 | 0.9328 | 16 | 0.7 | 6.7 | 6.0 | 47 | 62 | 4.11 | 1.46 |
| B3LYP/EPR-III | −0.2609 | 0.6516 | 0.9577 | 15 | 0.7 | 7.6 | 5.0 | 37 | 62 | 3.35 | 2.06 |
| (vii) $^{1}$H nucleus, all radicals | | | | | | | | | | | |
| PBE0/N07D | 0.5179 | 0.8936 | 0.9683 | 114 | 0.03 | 14.6 | 4.0 | 47 | 757 | 0.51 | 28.73 |
| B3LYP/6-31G* | 0.4130 | 0.9538 | 0.9714 | 114 | 0.01 | 15.0 | 3.6 | 31 | 186 | 0.46 | 32.74 |
| B3LYP/N07D | 0.3441 | 0.9613 | 0.9763 | 114 | 0.02 | 15.6 | 3.0 | 26 | 186 | 0.40 | 39.13 |
| B3LYP/TZVP | 0.2768 | 0.9180 | 0.9771 | 114 | 0.05 | 14.9 | 3.7 | 20 | 186 | 0.31 | 47.47 |
| B3LYP/EPR-III | 0.2991 | 0.9533 | 0.9857 | 107 | 0.1 | 16.5 | 2.4 | 23 | 233 | 0.33 | 49.99 |
| (viii) $^{1}$H nucleus, radicals **1**−**23** | | | | | | | | | | | |
| PBE0/N07D | 0.3940 | 1.0049 | 0.9795 | 74 | 0.1 | 14.0 | 1.6 | 56 | 757 | 0.46 | 30.33 |
| B3LYP/6-31G* | 0.2458 | 1.0648 | 0.9862 | 74 | 0.01 | 14.7 | 1.8 | 37 | 186 | 0.44 | 33.12 |
| B3LYP/N07D | 0.2149 | 1.0440 | 0.9876 | 74 | 0.02 | 14.5 | 1.6 | 30 | 186 | 0.37 | 39.06 |
| B3LYP/TZVP | 0.1607 | 0.9960 | 0.9874 | 74 | 0.05 | 14.1 | 1.2 | 24 | 186 | 0.25 | 56.14 |
| B3LYP/EPR-III | 0.2188 | 1.0081 | 0.9898 | 70 | 0.1 | 13.8 | 1.1 | 27 | 233 | 0.30 | 46.31 |
| (ix) $^{1}$H nucleus, radicals **24**−**38** | | | | | | | | | | | |
| PBE0/N07D | 0.6020 | 0.8315 | 0.9767 | 40 | 0.03 | 14.6 | 4.0 | 30 | 100 | 0.60 | 24.37 |
| B3LYP/6-31G* | 0.4654 | 0.8914 | 0.9742 | 40 | 0.1 | 15.0 | 3.6 | 21 | 60 | 0.48 | 30.75 |
| B3LYP/N07D | 0.3935 | 0.9142 | 0.9750 | 40 | 0.1 | 15.6 | 3.0 | 18 | 60 | 0.45 | 34.52 |
| B3LYP/TZVP | 0.3104 | 0.8743 | 0.9765 | 40 | 0.1 | 14.9 | 3.7 | 13 | 41 | 0.43 | 34.54 |
| B3LYP/EPR-III | 0.3212 | 0.9229 | 0.9857 | 37 | 0.1 | 16.5 | 2.4 | 14 | 53 | 0.39 | 42.17 |

[a] All calculations have been carried out on the geometries optimized at B3LYP/6-31G* level of theory. [b] E% (percent error) = $|a_{iso}(calc) - a_{iso}(exp)|/a_{iso}(exp)$. [c] MAD(Mean Absolute Deviation) = $1/N\sum_i^N |a_{iso}(calc) - a_{iso}(exp)|$.

methods considered herein. Both the 6-31G* and the N07D basis sets are equally appropriated for the theoretical evaluation of $^{14}$N hfccs of aromatic non nitroxide type radicals, but none of the five levels of theory are advisible for such a calculation in nitroxide radicals, given that all the ratios are very low. The thorough comparative analysis of all the parameters in Table 2 enables us to infer that, on balance, the most consistent results are provided by B3LYP/N07D level of theory, since it leads to overall predictions of hfccs of $^{14}$N and $^{1}$H nuclei of nitrogen aromatic radicals in a reasonable reliability. Unfortunately, the calculation of $a_{iso}(^{14}N)$ in the case of nitroxide radicals is particularly difficult, and all of the tested methodologies considerably underestimate the constants. The lack of reliability of these methods to compute nitrogen hfccs in nitroxide radicals has been previously outlined. A huge quantity of work has been

devoted to the theoretical investigation in depth on these species.[56−70] All of them, with remarkable recognition to the review by Improta and Barone,[56] point out that quantum mechanical methods that perform well for several classes of organic free radicals, do not lead to very good results for nitroxides, especially in the prediction of nitrogen isotropic hyperfine coupling. The disappointing results have been attributed to effects like vibrational averaging, conformational flexibility, and spin delocalization, which could play a significant role in determining this parameter, and are not correctly modeled by a static gas-phase QM approach. The accurate evaluation demands an integrated strategy able to provide a reliable description of the radical electronic structure. The first approach consisted of the correction of the DFT $a_{iso}(^{14}N)$ by means of a term computed with a post-Hartree−Fock method, such as quadratic configuration interaction (QCISD), coupled with purposely tailored basis sets.[56] Several studies have shown that this procedure leads to very good values of the nitrogen hfccs in nitroxides[62−65] but, unfortunately, it is not always applicable because of time-consuming computational difficulties. This is especially the case when both specific and bulk solvent effects, whose inclusion has proven to be non-negligible,[65,68] are properly accounted. More recently, a combination of discrete QM-continuum methods and MD simulations has been proposed as a more feasible integrated strategy, in which the averaged parameters are determined by running a molecular dynamics simulation of the implicit and/or explicit solvated radical and statistically sampling the resulting configurations.[57−60,62−67,69]

The deep analysis of the linear regression parameters calculated herein for $^{14}N$ nucleus of nitroxide species reveals that good correlations are obtained for the five fits in spite of the low range/MAD ratios. This implies that the theoretical values are far from the experimental ones but the deviation is always systematic, therefore, correctable as now it is known. B3LYP/EPR-III combination has the best correlation coefficient, but the value for the B3LYP/6-31G* fit is almost the same with the advantage of being much less computational demand and applicable to much more atoms than the former. Accordingly, we propose to use the B3LYP/6-31G* level of theory for the calculation of $a_{iso}(^{14}N)$ of nitroxide radicals and to scale the values by means of the linear fit obtained herein for such a calculation, that is:

$$a_N(\text{nitroxide}) = (a_N^{B3LYP/6-31G*} + 0.1105)/0.7536 \quad (1)$$

where both $a_N$ are given in Gauss. Figure 5 shows, as a noteworthy example, the result obtained by application of this equation to the nitroxide radicals studied in this work. The agreement between experimental data of nitrogen hfccs and the scaled-B3LYP/6-31G* values is excellent, with $R^2 > 0.95$, slope $\approx 1$, intercept $\approx 0$ and range/MAD $\approx 24$. Figure S2 in the SI displays the plots resulting from the application of a similar scale procedure to the other four levels of theory. All of them give very satisfactory results, but the range/MAD of the scaled-B3LYP/6-31G* fit is the highest, backing the selection of this computational protocol for this purpose.

Comparison of Figures 4 and 5 clearly indicates the improvement due to the addition of the corrective term

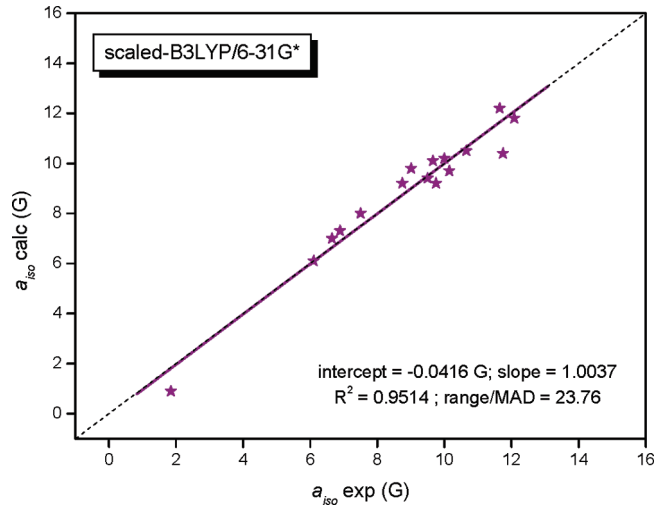



***Figure 5.*** Plot of theoretical vs experimental $a_{iso}$ for $^{14}N$ nuclei of radicals **24**−**38**, calculated by means of eq (1). The linear fit is represented by a solid line.

provided by the present work, allowing one to obtain reliable values of the $^{14}N$ hfccs of nitroxide species in a very simple way and with low computational cost.

## IV. Conclusions

An extensive study on the calculation of isotropic hyperfine coupling constants of aromatic radicals containing $^{14}N$ nucleus has been carried out by comparing 165 experimental hfccs to the corresponding data obtained from calculations at five different levels of theory: PBE0/N07D, B3LYP/6-31G*, B3LYP/N07D, B3LYP/TZVP, and B3LYP/EPR-III.

The results indicate that $a_{iso}$ of $^{1}H$ are predicted in very well agreement with the experimental values regardless of the level of theory employed, whereas significant differences are found in the case of the $^{14}N$ nucleus, being the selection of the basis set of fundamental importance, specially the number of components of *d* functions. In that regard, 6-31G* and N07D basis sets behave in a similar way in general, and predict hyperfine constants of $^{14}N$ closer to the experimental values than EPR-III and TZVP basis sets.

The thorough comparison of the results by a regression analysis points out that, on balance, the most consistent results are provided by the B3LYP/N07D level of theory, since it leads to overall predictions of hfccs of $^{14}N$ and $^{1}H$ nuclei of nitrogen aromatic radicals in a reasonable reliability.

A different tendency is observed for the calculated $a_{iso}(^{14}N)$ of non nitroxide and nitroxide type radicals. The first group is predicted quite accurately, whereas the values of the second type are considerably underestimated. As is widely known, the calculation of $a_{iso}(^{14}N)$ of nitroxide radicals is of particular complexity, and all the levels of theory investigated herein underestimate them in a great quantity. The N07D basis set combined with the PBE0 functional has shown to provide the most reliable values, albeit still important discrepancies are obtained.

Finally, a simple empirical method to obtain accurate values of the $a_{iso}(^{14}N)$ of nitroxide type radicals is provided, avoiding integrated methodologies which complexity and high computational cost restrict its application. This alterna-

tive consists on properly scaling the values obtained at B3LYP/6-31G* level of theory by means of the data provided by the linear fit corresponding to such a calculation.

**Acknowledgment.** The "Dirección General de Política Científica" of MEC (Spain) is gratefully acknowledged for financial support (MAT2008-06725-C03-02 and CTQ2010-19232).

**Supporting Information Available:** The name, the symmetry of the electronic ground state, and the total energies of the minimum of each radical computed at PBE0/N07D, B3LYP/6-31G*, B3LYP/N07D, B3LYP/TZVP, and B3LYP/EPR-III levels of theory (Table S1). Differences between the theoretical isotropic hyperfine coupling constants of $^{14}$N nuclei calculated at different levels of theory with 5 and 6 *d* functions, for a subset of ten radicals (Table S2). Plot of theoretical vs experimental $a_{iso}$ for $^{14}$N nuclei and $^{1}$H nuclei of the studied radicals, calculated at the five different levels of theory (Figure S1). Plot of theoretical vs experimental $a_{iso}$ for $^{14}$N nuclei of nitroxide radicals (**24**−**38**), calculated by properly scaling the data obtained at each level of theory (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Likhtenshtein, G. I. Nitroxide Spin Probes for Studies of Molecular Dynamics and Microstructure. In *Nitroxides: Applications in Chemistry, Biomedicine, and Materials Science*; Likhtenshtein, G. I.; Yamauchi, J., Nakatsuji, S., Smirnov, A. I., Tamura, R., Eds.; Wiley-VCH: Weinheim, Germany, 2008; pp 205−238.

(2) Gerson, F.; Huber, W. Electron-Nuclear Magnetic Interaction. In *Electron Spin Resonance Spectroscopy of Organic Radicals*; Wiley-VCH: Weinheim, Germany, 2003; pp 37−48.

(3) Corvaja, C. Introduction to Electron Paramagnetic Resonance. In *Electron Paramagnetic Resonance: A Practitioner'S Toolkit*; Brustolon, M., Giamello, E., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2009; pp 3−36.

(4) Kaupp, M.; Bühl, M.; Malkin, V. G. *Calculation of NMR and EPR Parameters: Theory and Applications*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2004. Munzarová, M. L. DFT Calculations of EPR hyperfine Coupling Tensors. In *Calculation of NMR and EPR Parameters: Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004; pp 463−482.

(5) Hermosilla, L.; Calle, P.; de la Vega, J. M. G.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 1114.

(6) Hermosilla, L.; Calle, P.; Sieiro, C. *Phosphorus Sulfur Silicon Relat. Elem.* **2005**, *180*, 1421.

(7) Hermosilla, L.; Calle, P.; de la Vega, J. M. G.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 7626.

(8) Hermosilla, L.; Calle, P.; de la Vega, J. M. G.; Sieiro, C. *J. Phys. Chem. A* **2006**, *110*, 13600.

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

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

(11) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.-Rev. Can. Chim.* **1992**, *70*, 560.

(12) Rega, N.; Cossi, M.; Barone, V. *J. Chem. Phys.* **1996**, *105*, 11060.

(13) Barone, V. *J. Phys. Chem.* **1995**, *99*, 11659.

(14) Hehre, W. J.; Ditchfie, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

(15) Harihara, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(16) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139.

(17) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751.

(18) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(19) *Gaussian 03, Revision E.01*; 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.

(20) Neta, P.; Fessenden, R. W. *J. Phys. Chem.* **1974**, *78*, 523.

(21) Nelsen, S. F.; Landis, R. T.; Kiehle, L. H.; Leung, T. H. *J. Am. Chem. Soc.* **1972**, *94*, 1610.

(22) Preston, K. F.; Sandall, J. P. B.; Sutcliffe, L. H. *Magn. Reson. Chem.* **1988**, *26*, 755.

(23) Mukai, K.; Nishiguchi, H.; Ishizu, K.; Deguchi, Y.; Takaki, H. *Bull. Chem. Soc. Jpn.* **1967**, *40*, 2731.

(24) Neugebauer, F. A.; Bamberger, S. *Chem. Ber.* **1974**, *107*, 2362.

(25) Neugebauer, F. A.; Fischer, H.; Bamberger, S.; Smith, H. O. *Chem. Ber.* **1972**, *105*, 2694.

(26) Scheffler, K.; Stegmann, H. B. *Tetrahedron Lett.* **1968**, *9*, 3619.

(27) Clarke, D.; Gilbert, B. C.; Hanson, P. *J. Chem. Soc., Perkin Trans 2* **1977**, *4*, 517.

(28) Samunl, A.; Neta, P. *J. Phys. Chem.* **1973**, *77*, 1629.

(29) Itoh, M.; Nagakura, S. *Tetrahedron Lett.* **1965**, *6*, 417.

(30) Hermolin, J.; Levin, M.; Ikegami, Y.; Sawayanagi, M.; Kosower, E. M. *J. Am. Chem. Soc.* **1981**, *103*, 4795.

(31) Braun, D.; Peschk, G.; Hechler, E. *Chemiker-Ztg.* **1970**, *94*, 703.

(32) Hyde, J. S.; Robert, C.; Sneed, J.; Rist, G. H. *J. Chem. Phys.* **1969**, *51*, 1404.

(33) Hudson, R. F.; Lawson, A. J.; Record, K. A. F. *J. Chem. Soc., Chem. Commun.* **1974**, *12*, 488.

(34) Dalal, N. S.; Rippmeester, J. A.; Reddoch, A. H. *J. Magn. Reson.* **1978**, *31*, 471.

(35) Biehl, R.; Moebius, K.; O'Connor, S. E.; Walter, R. I.; Zimmermann, H. *J. Phys. Chem.* **1979**, *83*, 3449.

(36) Dalal, N. S.; Kennedy, D. E.; McDowell, C. A. *J. Chem. Phys.* **1974**, *61*, 1689.

(37) Neugebauer, F. A. *Tetrahedron* **1970**, *26*, 4843.

(38) Neugebauer, F. A.; Russell, G. A. *J. Org. Chem.* **1968**, *33*, 2744.

(39) Neugebauer, F. A. *Chem. Ber.* **1969**, *102*, 1339.

(40) Kuhn, R.; Trischmann, H. *Mh. Chem.* **1964**, *95*, 457.

(41) Brunner, H.; Hausser, K. H.; Neugebauer, F. A. *Tetrahedron* **1971**, *27*, 3611.

(42) Neugebauer, F. A. *Tetrahedron* **1970**, *26*, 4853.

(43) Mukai, K.; Yamamoto, T.; Kohno, M.; Azuma, N.; Ishizu, K. *Bull. Chem. Soc. Jpn.* **1974**, *47*, 1797.

(44) Kopf, P.; Morokuma, K.; Kreilick, R. *J. Chem. Phys.* **1971**, *54*, 105.

(45) Barbarella, G.; Rassat, A. *Bull. Soc. Chim. Fr.* **1969**, 2378.

(46) Terabe, S.; Kuruma, K.; Konaka, R. *J. Chem. Soc., Perkin. Trans. 2* **1973**, *9*, 1252.

(47) Nishikawa, T.; Someno, K. *Bull. Chem. Soc. Jpn.* **1974**, *47*, 2881.

(48) Briere, R.; Chapelet-Letourneux, G.; Lemaire, H.; Rassat, A. *Mol. Phys.* **1971**, *20*, 211.

(49) Ishizu, K.; Nagai, H.; Mukai, K.; Kohno, M.; Yamamoto, T. *Chem. Lett.* **1973**, *2*, 1261.

(50) Bruni, P.; Greci, L. *J. Heterocyclic Chem.* **1972**, *9*, 1455.

(51) Fischer, P. H. H.; Neugebauer, F. A. *Z. Naturforsch.* **1964**, *19a*, 1514.

(52) Aurich, H. G.; Hahn, K.; Stork, K.; Weiss, W. *Tetrahedron* **1977**, *33*, 969.

(53) Chiu, M. F.; Gilbert, B. C.; Hanson, P. *J. Chem. Soc. B* **1970**, 1700.

(54) Aurich, H. G.; Hahn, K.; Stork, K. *Angew. Chem.-Int. Ed. Engl.* **1975**, *14*, 551.

(55) Ramsbottom, J. V.; Waters, W. A. *J. Chem. Soc. B* **1966**, 132.

(56) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231.

(57) Pavone, M.; Benzi, C.; De Angelis, F.; Barone, V. *Chem. Phys. Lett.* **2004**, *395*, 120.

(58) Houriez, C.; Ferre, N.; Masella, M.; Siri, D. *J. Chem. Phys.* **2008**, *128*, 244504.

(59) Houriez, C.; Ferre, N.; Masella, M.; Siri, D. *Theochem.-J. Mol. Struct.* **2009**, *898*, 49.

(60) Cimino, P.; Pedone, A.; Stendardo, E.; Barone, V. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3741.

(61) Rajca, A.; Vale, M.; Rajca, S. *J. Am. Chem. Soc.* **2008**, *130*, 9099.

(62) Tedeschi, A. M.; D'Errico, G.; Busi, E.; Basosi, R.; Barone, V. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2180.

(63) Improta, R.; Scalmani, G.; Barone, V. *Chem. Phys. Lett.* **2001**, *336*, 349.

(64) Improta, R.; di Matteo, A.; Barone, V. *Theor. Chem. Acc.* **2000**, *104*, 273.

(65) Saracino, G. A. A.; Tedeschi, A.; D'Errico, G.; Improta, R.; Franco, L.; Ruzzi, M.; Corvaia, C.; Barone, V. *J. Phys. Chem. A* **2002**, *106*, 10700.

(66) Cimino, P.; Pavone, M.; Barone, V. *J. Phys. Chem. A* **2007**, *111*, 8482.

(67) Pavone, M.; Cimino, P.; Crescenzi, O.; Sillanpaa, A.; Barone, V. *J. Phys. Chem. B* **2007**, *111*, 8928.

(68) Owenius, R.; Engstrom, M.; Lindgren, M.; Huber, M. *J. Phys. Chem. A* **2001**, *105*, 10967.

(69) Barone, V. *Chem. Phys. Lett.* **1996**, *262*, 201.

(70) Ikryannikova, L. N.; Ustynyuk, L. Y.; Tikhonov, A. N. *Magn. Reson. Chem.* **2010**, *48*, 337.

CT1006136

JCTC Journal of Chemical Theory and Computation

# Oscillator Strengths in ONIOM Excited State Calculations

Marco Caricato,*[,†] Thom Vreven,[‡] Gary W. Trucks,[†] and Michael J. Frisch[†]

*Gaussian, Inc., 340 Quinnipiac St., Bldg. 40, Wallingford, Connecticut 06492, United States, and Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, United States*

Received November 2, 2010

**Abstract:** We compute oscillator strengths with the ONIOM (Our own N-layer Integrated molecular Orbital molecular Mechanics) hybrid method between ground and valence excited states and compare the results with the high level of theory equation of motion coupled cluster singles and doubles (EOM-CCSD). This work follows our previous studies in which we validated the ability of ONIOM to compute accurate transition energies compared to EOM-CCSD. We test various levels of theory and molecular systems, as well as the effect of the link atom bond length. Our results show that oscillator strengths can be accurately computed with ONIOM, provided that a sensible choice of the partitioning and of the low level method is made. Being able to calculate both the transition energy and the oscillator strength, ONIOM represents a promising approach to completely characterize valence excited states of molecules that are too large to be studied with a conventional high-accuracy method.

## 1. Introduction

Methods that combine two (or more) levels of theory are now widely used to study ground state phenomena and processes of large systems.[1] Such "hybrid methods" divide the system into a region of major interest treated at a high level of theory, while the rest (substituent effect) is treated at a lower and less computationally demanding level. Most of these methods combine a quantum mechanical level with a molecular mechanical level, QM/MM,[1−4] and the energy is expressed as a summation. On the other hand, the ONIOM (Our own N-layer Integrated molecular Orbital molecular Mechanics)[5−15] energy is formulated as extrapolation. Therefore, it can easily combine more than two computational levels as well as integrate two or more different quantum mechanical levels, QM/QM. In ONIOM, open valencies in the *model* system generated from cutting covalent bonding between regions are saturated with link atoms (typically hydrogens). A link atom is placed in the same direction of the atom it replaces, with a distance scaled by a factor proportional to the ratio between the original bond length and the typical bond distance of the atoms involved.

While the use of hybrid methods for ground state problems has been well documented, their use in excited state calculation is still limited. This is especially so when two QM levels of theory are combined.[1] In two recent studies,[16,17] we investigated the ability of the ONIOM hybrid method to reproduce equation of motion coupled cluster singles and doubles (EOM-CCSD)[18−24] vertical excitation energies. The EOM-CCSD results on the entire system (the *target*) were used as a reference, and the performance of ONIOM was compared either to conventional calculations with a lower level of theory on the entire system or to EOM-CCSD only on the core region. We analyzed the effect of the partitioning, of the choice of the low level methods and basis sets, and of the link atom bond length. We found that, provided certain guidelines for the partitioning[15,17] are followed, ONIOM is able to accurately approximate EOM-CCSD while providing a drastic reduction in computational time. ONIOM with time-dependent density functional theory (TDDFT) as the low level provided the best overall performance. Also, the exact value of the link atom bond length did not significantly influence the ONIOM results, analogous to what was found for ground state calculations;[25] hence, the same definition for the link atom bond length used for the ground state can

* To whom correspondence should be addressed. E-mail: marco@gaussian.com.

† Gaussian, Inc.

‡ University of Massachusetts.

10.1021/ct1006289 
Published on Web 12/21/2010

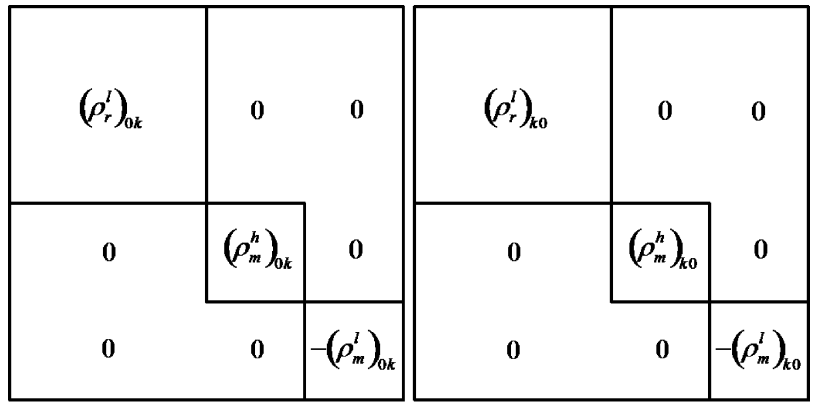



***Figure 1.*** Integrated ONIOM transition densities.

be retained for excited state calculations. Finally, we also proposed several new guidelines for using ONIOM in excited states studies.[17]

Although the energy is the primary quantity that needs to be considered to assess the performance of ONIOM, transition properties are also important for the investigation of electronic excitations. In this paper, we focus on the oscillator strength since this quantity is directly related to the intensity of a transition. Our comparison will be with this property computed for the entire system at the *high* level of theory, because this is the reference value that we aim to reproduce (*target*). We consider two molecular systems from our previous works and two new systems with large oscillator strengths. For this paper, as in our previous studies, we limit our investigation to valence states. The *model* system for each molecule is chosen according to the ONIOM guidelines for partitioning[15] and based on the results of our previous studies.[16,17] For this *model* system, the ONIOM *high* level calculation is performed at the *target* level of theory while various levels of theory are tested for the ONIOM *low* level calculations. Only QM/QM combinations are considered in this work, as in refs 16 and 17, since a MM method in the *low* level does not directly contribute to the excited state calculation.

The results in this paper show that transition properties can be accurately computed with ONIOM, and that the same considerations employed when computing the transition energy apply to the oscillator strength.

The paper is organized as follows. The formulas for the calculation of ONIOM transition energies and properties are reported in section 2. Computational details are given in section 3. The results of the calculations are presented in section 4, while section 5 contains a discussion of these results and concluding remarks.

## 2. Theory

The ONIOM energy for a two-layer system with mechanical embedding is written as an extrapolation:

$$E^{\mathrm{ONIOM}} = E^{\mathrm{high}}_{\mathrm{model}} + E^{\mathrm{low}}_{\mathrm{real}} - E^{\mathrm{low}}_{\mathrm{model}} \tag{1}$$

where *real* and *model* refer to the full system and to the core region, respectively. The transition energy ($\Delta E$) in the ONIOM scheme can be expressed as the difference of the ONIOM energies of the $k$th and the ground states:

$$\begin{aligned}\Delta E^{\mathrm{ONIOM}} &= E^{k,\mathrm{ONIOM}} - E^{\mathrm{ONIOM}} \\ &= (E^{k,\mathrm{high}}_{\mathrm{model}} + E^{k,\mathrm{low}}_{\mathrm{real}} - E^{k,\mathrm{low}}_{\mathrm{model}}) - (E^{\mathrm{high}}_{\mathrm{model}} + E^{\mathrm{low}}_{\mathrm{real}} - E^{\mathrm{low}}_{\mathrm{model}}) \\ &= (E^{k,\mathrm{high}}_{\mathrm{model}} - E^{\mathrm{high}}_{\mathrm{model}}) + (E^{k,\mathrm{low}}_{\mathrm{real}} - E^{\mathrm{low}}_{\mathrm{real}}) - (E^{k,\mathrm{low}}_{\mathrm{model}} - E^{\mathrm{low}}_{\mathrm{model}}) \\ &= \Delta E^{\mathrm{high}}_{\mathrm{model}} + \Delta E^{\mathrm{low}}_{\mathrm{real}} - \Delta E^{\mathrm{low}}_{\mathrm{model}}\end{aligned} \tag{2}$$

The ONIOM oscillator strength, $f^{\mathrm{ONIOM}}$, is calculated as

$$f^{\mathrm{ONIOM}} = \frac{2}{3}\Delta E^{\mathrm{ONIOM}} D^{\mathrm{ONIOM}} \tag{3}$$

where $D^{\mathrm{ONIOM}}$ is the ONIOM dipole strength defined as

$$D^{\mathrm{ONIOM}} = \sum_i (\mu_i^{\mathrm{ONIOM}})_{0k} \times (\mu_i^{\mathrm{ONIOM}})_{k0} \quad i = x, y, z \tag{4}$$

and $(\mu_i^{\mathrm{ONIOM}})_{0k}$ are the components of the integrated ONIOM transition dipole:

$$(\mu_i^{\mathrm{ONIOM}})_{0k} = Tr[(\rho_m^h)_{0k}(\mu_m^h)_i] + Tr[(\rho_r^l)_{0k}(\mu_r^l)_i] - Tr[(\rho_m^l)_{0k}(\mu_m^l)_i] \tag{5}$$

where $\mu^{\mathrm{h/l}}_{\mathrm{m/r}}$ are the dipole integrals of the various subcalculations (h/l indicates the *high* or *low* level of theory, and m/r indicates the *model* or *real* system). A similar expression is used to compute $(\mu_i^{\mathrm{ONIOM}})_{k0}$. In order to compute transition properties, integrated ONIOM transition densities must be defined, $\rho_{0k}^{\mathrm{ONIOM}}$ and $\rho_{k0}^{\mathrm{ONIOM}}$. These are shown in Figure 1, where the ground to excited state ($\rho_{0k}^{\mathrm{ONIOM}}$) and the excited to ground state ($\rho_{k0}^{\mathrm{ONIOM}}$) densities are represented in terms of the transition densities of the subcalculations. Two different integrated transition density matrices are necessary when a method like EOM-CCSD is used since this method has a non-Hermitian Hamiltonian.[20] However, $\rho_{0k} = \rho_{k0}$ for methods like configuration interaction singles (CIS), time-dependent Hartree−Fock (TDHF), and TDDFT.

Alternatively, a simpler formula may be used to compute $f^{\mathrm{ONIOM}}$ by directly integrating the oscillator strengths of the subcalculations:

$$\tilde{f}^{\mathrm{ONIOM}} = f^{\mathrm{high}}_{\mathrm{model}} + f^{\mathrm{low}}_{\mathrm{real}} - f^{\mathrm{low}}_{\mathrm{model}} \tag{6}$$

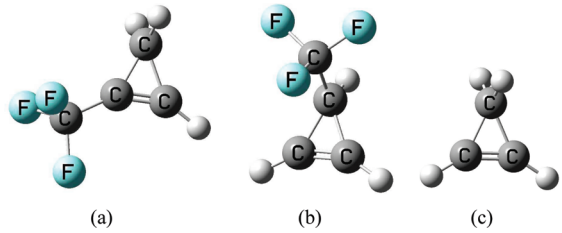



***Figure 2.*** Structures of 1-trifluoro-methyl-cyclopropene (a), 3-trifluoro-methyl-cyclopropene (b), and the *model* system (c).

which does not require the integration of the transition dipole. $\tilde{f}^{\mathrm{ONIOM}}$ is an approximation of $f^{\mathrm{ONIOM}}$ in eq 3, as it is easy to verify that $f^{\mathrm{ONIOM}} = \tilde{f}^{\mathrm{ONIOM}} + \textit{cross terms}$. However, since the subcalculation transition dipoles are needed for both $f^{\mathrm{ONIOM}}$ and $\tilde{f}^{\mathrm{ONIOM}}$, there is not a particular computational advantage in using the latter. Therefore, we will not further consider $\tilde{f}^{\mathrm{ONIOM}}$.

## 3. Computational Details

The ground state geometries of the entire molecules are optimized at the B3LYP level with the 6-311+G** basis set and subsequently used for all the excited state calculations. The *target* oscillator strength and transition energy (i.e., that we want to reproduce with ONIOM), $f^{\mathrm{target}}$ and $\Delta E^{\mathrm{target}}$, respectively, are computed at the EOM-CCSD/6-311+G** level of theory on the entire (*real*) system. For ONIOM, only one level of theory is used in the *high* level calculation on the *model* system: EOM-CCSD/6-311+G** (i.e., the same as the *target*). For the *low* level calculations (on the *model* and *real* systems), we test CIS, TDHF, and TDDFT (with the B3LYP functional[26−28]) with the 6-311+G** basis set. We also consider EOM-CCSD/6-31+G* as a *low* level method. In the following, we refer to the 6-311+G** and 6-31+G* basis sets as "L" and "M", respectively, for consistency with refs 16 and 17. For the ketene, we also test three functionals other than B3LYP, namely, CAM-B3LYP,[29,30] BLYP,[31,32] and LC-BLYP.[31−34] All of the calculations are performed with the Gaussian 09 suite of programs.[35] In summary, we consider the following ONIOM combinations:

- ONIOM(EOM/L:EOM/M)
- ONIOM(EOM/L:TDDFT/L)
- ONIOM(EOM/L:TDHF/L)
- ONIOM(EOM/L:CIS/L)

In the next section, these ONIOM combinations are compared with the conventional calculations at the corresponding *low* level of theory on the entire system ($f^{\mathrm{low}}_{\mathrm{real}}$ and $\Delta E^{\mathrm{low}}_{\mathrm{real}}$), and with the *high* level of theory on the *model* system ($f^{\mathrm{high}}_{\mathrm{model}}$ and $\Delta E^{\mathrm{high}}_{\mathrm{model}}$). In fact, $f^{\mathrm{low}}_{\mathrm{real}}$ and $f^{\mathrm{high}}_{\mathrm{model}}$ can be considered as approximations of $f^{\mathrm{target}}$ when the latter is computationally too expensive to evaluate (equivalently $\Delta E^{\mathrm{low}}_{\mathrm{real}}$ and $\Delta E^{\mathrm{high}}_{\mathrm{model}}$ for $\Delta E^{\mathrm{target}}$). Therefore, it is interesting to test ONIOM versus these approximated conventional calculations.

## 4. Results

The ability of ONIOM to reproduce the *target* oscillator strengths is tested on four molecular systems, shown in Figures 2−4 with their respective *model* systems. Although

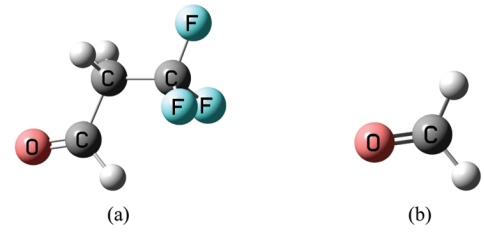



***Figure 3.*** Structures of 3,3,3-trifluoropropanal (a) and the *model* system (b).

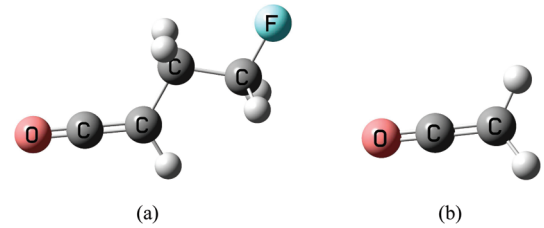



***Figure 4.*** Structures of 4-fluorobut-1-en-1-one (a) and the *model* system (b).

the number of test cases in this work is limited, they are analyzed in great detail, and a large variety of possible sources of errors is considered. Therefore, even though the results presented in this section are preliminary, they provide useful insight on the ONIOM accuracy for this property. The first two systems, the substituted cyclopropenes, appeared in our previous works where we analyzed the dependence of $\Delta E^{\mathrm{ONIOM}}$ on the choice of the *model* system, the *low* level of theory, and the link atom bond length.[16,17] These systems represent good candidates to study the ONIOM performance on transition properties because we already know how the various choices of the subcalculations influence the transition energy. For this reason, we only consider the *model* systems that performed best in refs 16 and 17, which indeed follow the ONIOM guidelines for the partitioning.[15] We consider the second transition for both systems since it has a larger oscillator strength. Unfortunately, we could not select other systems from our previous studies because $f^{\mathrm{target}}$ was negligible and thus not suitable for this investigation. The two new systems chosen for this study are 3,3,3-trifluoropropanal and 4-fluorobut-1-en-1-one (in the following, we will refer to these systems as the "aldehyde" and the "ketene", respectively), for which we examine two bright states. Both molecules are forced in the $C_s$ geometry. All of the transitions considered are $\pi \rightarrow \pi^*$. In addition to the oscillator strength, we also discuss the ONIOM performance on $\Delta E$ in order to understand how their relative performances are related.

For the cyclopropenes, the transition energies are reported in Table 1 and the oscillator strengths in Table 2. An interesting aspect of these two molecules is that the same *model* system can be used for both (albeit with some geometrical differences due to the different *real* systems' geometries). As discussed in ref 16, $\Delta E^{\mathrm{ONIOM}}$ is a very good estimates of $\Delta E^{\mathrm{target}}$ for both molecules, especially with EOM/M and TDDFT/L in the *low* level. Although $\Delta E^{\mathrm{high}}_{\mathrm{model}}$ for the first system is very close to the *target*, ONIOM(EOM/L:EOM/M) is even closer, while the error with ONIOM(EOM/L:TDDFT/L) is below 0.1 eV. For the second system,

***Table 1.*** Transition Energies (eV) for the Substituted Cyclopropenes

| | 1-trifluoro-methyl-cyclopropene | | | 3-trifluoro-methyl-cyclopropene | | |
|---|---|---|---|---|---|---|
| X | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ |
| EOM/L | 7.09[d] | 7.10[e] | | 7.48[d] | 7.04[e] | |
| EOM/M | 7.15 | 7.15 | 7.09 | 7.57 | 7.10 | 7.52 |
| TDDFT/L | 6.19 | 6.27 | 7.02 | 6.63 | 6.20 | 7.48 |
| TDHF/L | 6.67 | 6.58 | 7.18 | 6.98 | 6.57 | 7.46 |
| CIS/L | 6.99 | 6.86 | 7.22 | 7.29 | 6.85 | 7.49 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $\Delta E^{target}$: EOM/L level on the *real* system. [e] $\Delta E^{high}_{model}$: EOM/L level on the *model* system.

***Table 2.*** Oscillator Strengths for the Substituted Cyclopropenes

| | 1-trifluoro-methyl-cyclopropene | | | 3-trifluoro-methyl-cyclopropene | | |
|---|---|---|---|---|---|---|
| X | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ |
| EOM/L | 0.1128[d] | 0.0792[e] | | 0.0810[d] | 0.0783[e] | |
| EOM/M | 0.1124 | 0.0794 | 0.1120 | 0.0800 | 0.0784 | 0.0799 |
| TDDFT/L | 0.1020 | 0.0702 | 0.1155 | 0.0671 | 0.0677 | 0.0770 |
| TDHF/L | 0.1474 | 0.1202 | 0.1022 | 0.1209 | 0.1205 | 0.0772 |
| CIS/L | 0.1748 | 0.1493 | 0.0977 | 0.1537 | 0.1500 | 0.0789 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $f^{target}$: EOM/L level on the *real* system. [e] $f^{high}_{model}$: EOM/L level on the *model* system.

***Table 3.*** Transition Energies (eV) and Oscillator Strengths for the Aldehyde

| X | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ |
|---|---|---|---|---|---|---|
| EOM/L | 9.36[d] | 10.00[e] | | 0.1423[f] | 0.1577[g] | |
| EOM/M | 9.41 | 10.13 | 9.28 | 0.1369 | 0.1854 | 0.1161 |
| TDDFT/L | 8.60 | 9.50 | 9.10 | 0.1171 | 0.1277 | 0.1416 |
| TDHF/L | 9.28 | 9.38 | 9.90 | 0.2391 | 0.2422 | 0.1582 |
| CIS/L | 9.66 | 9.74 | 9.92 | 0.2628 | 0.2824 | 0.1439 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $\Delta E^{target}$: EOM/L level on the *real* system. [e] $\Delta E^{high}_{model}$: EOM/L level on the *model* system. [f] $f^{target}$: EOM/L level on the *real* system. [g] $f^{high}_{model}$: EOM/L level on the *model* system.

ONIOM provides a large improvement over $\Delta E^{high}_{model}$ and $\Delta E^{low}_{real}$ for all of the *low* level methods. Table 2 shows that the same good performance is shared by the oscillator strength calculations. In this case, the first molecule shows the larger improvement by using ONIOM instead of the conventional calculations ($f^{high}_{model}$ and $f^{low}_{real}$) to approximate the *target*, except for EOM/M where $f^{low}_{real}$ is very close to $f^{target}$; nevertheless, the ONIOM results are very close to the *target* even for EOM/M. For the second molecule, $f^{high}_{model}$ is already very close to $f^{target}$, but the ONIOM results also have very small errors. Additionally, ONIOM always improves over $f^{low}_{real}$.

The results for the aldehyde are reported in Table 3. The performances of ONIOM(EOM/L:EOM/M) and ONIOM-(EOM/L:TDDFT/L) on the transition energy are very good. The former maintains an already small $\Delta E^{low}_{real}$ error, and the latter improves on both $\Delta E^{high}_{model}$ and $\Delta E^{low}_{real}$. This is not the case with CIS/L and TDHF/L in the *low* level. For these two methods, the improvement over $\Delta E^{high}_{model}$ is in the right direction, but it is not enough to provide a good estimate of $\Delta E^{target}$. The reason is probably the lack of correlation effects, which are important for obtaining a reliable description of the substituent effect. $f^{target}$ is well reproduced with ONIOM and TDDFT/L, TDHF/L, and CIS/L as low methods. The poor performance of ONIOM(EOM/L:EOM/M) for the

***Table 4.*** Basis Set Dependence of the Transition Energies (eV) and Oscillator Strengths for the Aldehyde

| X | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ |
|---|---|---|---|---|---|---|
| EOM/L | 9.36[d] | 10.00[e] | | 0.1423[f] | 0.1577[g] | |
| EOM/M | 9.41 | 10.13 | 9.28 | 0.1369 | 0.1854 | 0.1161 |
| CIS/L | 9.66 | 9.74 | 9.92 | 0.2628 | 0.2824 | 0.1439 |
| CIS/M | 9.71 | 9.84 | 9.87 | 0.2623 | 0.3097 | 0.1253 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $\Delta E^{target}$: EOM/L level on the *real* system. [e] $\Delta E^{high}_{model}$: EOM/L level on the *model* system. [f] $f^{target}$: EOM/L level on the *real* system. [g] $f^{high}_{model}$: EOM/L level on the *model* system.

***Table 5.*** Transition Energies (eV) and Oscillator Strengths for the Ketene

| X | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ |
|---|---|---|---|---|---|---|
| EOM/L | 10.09[d] | 10.74[e] | | 0.7675[f] | 0.9801[g] | |
| EOM/M | 10.16 | 10.82 | 10.07 | 0.6392 | 0.9825 | 0.6355 |
| TDDFT/L | 9.31 | 10.37 | 9.67 | 0.3473 | 0.8211 | 0.4385 |
| TDHF/L | 10.50 | 10.81 | 10.42 | 0.8353 | 0.7804 | 1.0336 |
| CIS/L | 10.87 | 11.18 | 10.42 | 0.5967 | 0.9768 | 0.6034 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $\Delta E^{target}$: EOM/L level on the *real* system. [e] $\Delta E^{high}_{model}$: EOM/L level on the *model* system. [f] $f^{target}$: EOM/L level on the *real* system. [g] $f^{high}_{model}$: EOM/L level on the *model* system.

oscillator strength can be explained by the inadequacy of the "M" basis set for the *model* system calculation. Indeed, repeating the calculation at the CIS level with this basis set (reported in Table 4), we find that the trends EOM/L → EOM/M and CIS/L → CIS/M are similar. Although $f^{ONIOM}$ with EOM/M in the *low* level is not unreasonable, it certainly increases the error compared to $f^{high}_{model}$ and $f^{low}_{real}$.

The last system is a ketene for which we consider a very bright transition, $f^{target} = 0.7675$. The results are reported in Table 5. ONIOM improves the agreement with $\Delta E^{target}$ with all the *low* level methods compared to $\Delta E^{high}_{model}$ and $\Delta E^{low}_{real}$ (for EOM/M, the improvement is small, as $\Delta E^{low}_{real}$ is already

**Table 6.** Functional Dependence of the Transition Energies (eV) and Oscillator Strengths for the Ketene

| X | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $\Delta E^{ONIOM,c}$ | $f_{real}$[a] | $f_{model}$[b] | $f^{ONIOM,c}$ |
|---|---|---|---|---|---|---|
| EOM/L | 10.09[d] | 10.74[e] | | 0.7675[f] | 0.9801[g] | |
| B3LYP/L | 9.31 | 10.37 | 9.67 | 0.3473 | 0.8211 | 0.4385 |
| CAM-B3LYP/L | 9.75 | 10.49 | 9.99 | 0.6355 | 0.8399 | 0.7448 |
| BLYP/L | 9.80 | 10.09 | 10.45 | 0.1634 | 0.7500 | 0.2581 |
| LC-BLYP/L | 10.18 | 10.67 | 10.25 | 0.8521 | 0.8784 | 0.9480 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system. [c] ONIOM(EOM/L:X), where X is the level of theory specified in the first column. [d] $\Delta E^{target}$: EOM/L level on the *real* system. [e] $\Delta E^{high}_{model}$: EOM/L level on the *model* system. [f] $f^{target}$: EOM/L level on the *real* system. [g] $f^{high}_{model}$: EOM/L level on the *model* system.

**Table 7.** Transition Energies (eV) and Oscillator Strengths for the Ketene with the daug-cc-pVTZ Basis Set

| | $\Delta E_{real}$[a] | $\Delta E_{model}$[b] | $f_{real}$[a] | $f_{model}$[b] |
|---|---|---|---|---|
| EOM | | 10.48 | | 0.7088 |
| B3LYP | 9.17 | 10.25 | 0.3246 | 0.4940 |
| TDHF | 10.35 | 10.64 | 0.5330 | 0.5803 |
| CIS | 10.69 | 10.91 | 0.6253 | 0.6669 |
| CAM-B3LYP | 9.55 | 10.18 | 0.3101 | 0.5930 |
| BLYP | 9.67 | 10.75 | 0.1607 | 0.4010 |
| LC-BLYP | 10.07 | 10.52 | 0.6095 | 0.7261 |

[a] Conventional calculation on the *real* system. [b] Conventional calculation on the *model* system.

very close to $\Delta E^{target}$). For the oscillator strength, the ONIOM performance varies depending on the *low* level method. $f^{ONIOM}$ with EOM/M is very close to $f^{target}$. ONIOM(EOM/L:TDDFT/L) and ONIOM(EOM/L:CIS/L) move $f^{high}_{model}$ and $f^{low}_{real}$ in the right direction; although for TDDFT/L, $f^{high}_{model}$ is closer to the *target* than $f^{ONIOM}$. ONIOM(EOM/L:TDHF/L) is very close to $f^{high}_{model}$, but the correction goes in the wrong direction. The poor performance of TDDFT/L as a *low* level in this case, contrary to the previous ones, is related to the choice of the particular functional: B3LYP does not satisfactorily reproduce the long-range effect of the substituent group. This can be recovered by employing a functional that better describes this effect, such as CAM-B3LYP. ONIOM(EOM/L:CAM-B3LYP/L) provides results in very good agreement with the *target* calculation, both for the energy and the oscillator strength, see Table 6. The lack of a correct description of the long-range effect in B3LYP as a cause of its poor performance is confirmed by considering the results obtained with the BLYP functional with and without long-range corrections, also in Table 6. Although BLYP is not a good choice for computing excited state energy and properties with ONIOM and in conventional calculations,[36] the trend between B3LYP and BLYP when long-range corrections are introduced is similar. This problem could be partly overcome by considering a much larger basis set, and in Table 7 we report the data computed with the daug-cc-pVTZ basis, which has a double set of diffuse functions. This basis set is too large for performing the calculation on the entire system at the EOM-CCSD level, but the results in Table 7 show that the difference between B3LYP and CAM-B3LYP (and BLYP and LC-BLYP) is greatly reduced.

Finally, we examine the effect of the link atom bond length ($r_{LA}$) on the oscillator strength by computing it when moving away from the standard $r_{LA}$ value ($r^0_{LA}$) in a range of ±0.3 Å. We only consider one *low* level method for the sake of clarity (we choose TDDFT since it provides on average the best performance). Figure 5 reports the variation of $f^{ONIOM}$ as well as $f^{high}_{model}$ and $f^{low}_{model}$ as a function of the $r_{LA}$ shift for the substituted cyclopropenes: $r^0_{LA} = 1.067$ Å and $= 1.092$ Å, respectively. The figures show that ONIOM only has a small dependence on $r_{LA}$ for the oscillator strength. In Figure 5a, the variation is also small for the subcalculations (in the range of ±0.1 Å), whereas in Figure 5b, the *model* system calculations have the same dependence on $r_{LA}$. Thus, $f^{ONIOM}$ benefits from the error cancellation. These results are similar to the behavior of $\Delta E^{ONIOM}$ for the same systems (compare with Figures 19 and 20 in ref 17). For the aldehyde, we consider the shift of the transition energy ($\Delta E^{high}_{model}$, $\Delta E^{low}_{model}$ and $\Delta E^{ONIOM}$) with $\Delta r_{LA}$ with respect to $r^0_{LA}$, and the dependence of the oscillator strength ($f^{high}_{model}$, $f^{low}_{model}$, and $f^{ONIOM}$) on $\Delta r_{LA}$,

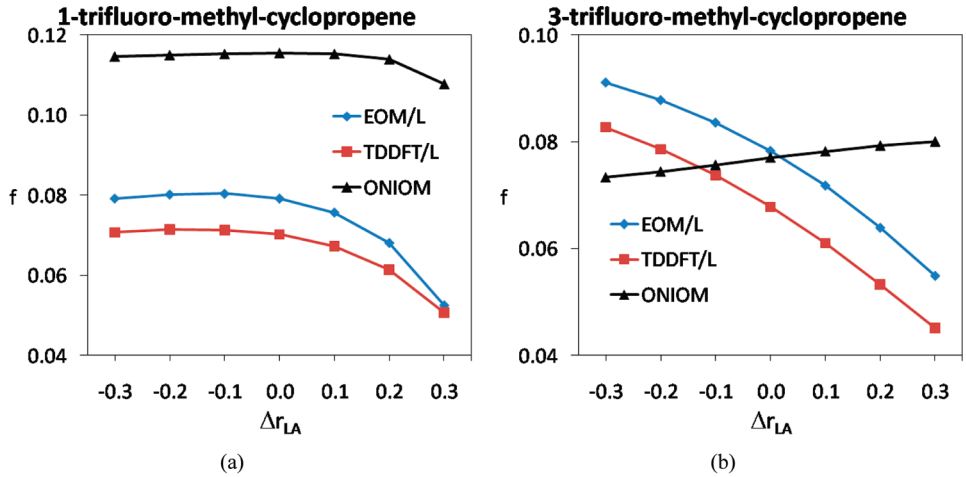



**Figure 5.** $f^{high}_{model}$, $f^{low}_{model}$, and $f^{ONIOM}$ dependence on $\Delta r_{LA}$ (Å) for the substituted cyclopropenes.

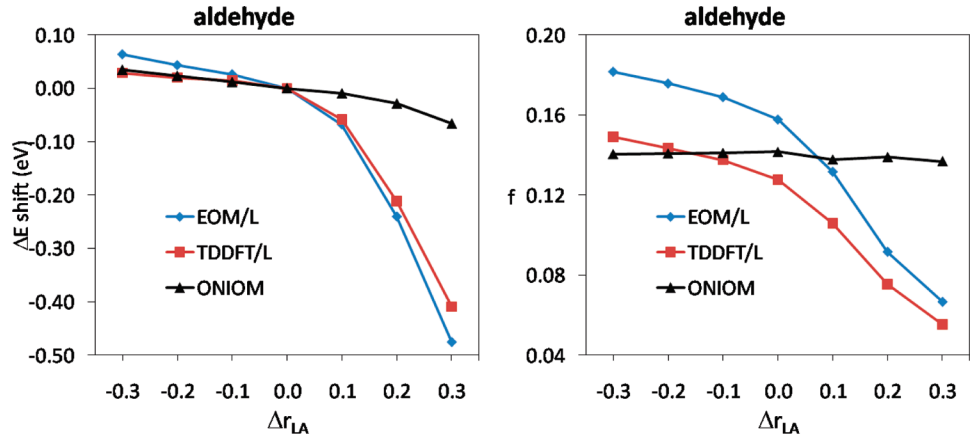



***Figure 6.*** (a) $\Delta E^{high}_{model}$, $\Delta E^{low}_{model}$, and $\Delta E^{ONIOM}$ shift from $r^0_{LA}$. (b) $f^{high}_{model}$, $f^{low}_{model}$, and $f^{ONIOM}$ dependence on $\Delta r_{LA}$ (Å) for the aldehyde.

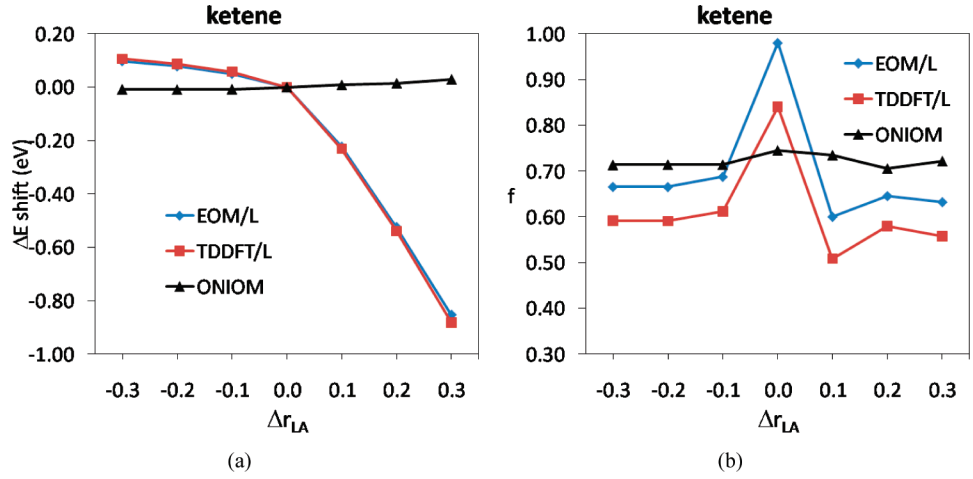



***Figure 7.*** (a) $\Delta E^{high}_{model}$, $\Delta E^{low}_{model}$, and $\Delta E^{ONIOM}$ shift from $r^0_{LA}$. (b) $f^{high}_{model}$, $f^{low}_{model}$, and $f^{ONIOM}$ dependence on $\Delta r_{LA}$ (Å) for the ketene. TDDFT in this case refers to the CAM-B3LYP functional.

shown in Figure 6. The precise value of $r_{LA}$ ($r^0_{LA}$ = 1.105 Å; the other C−H bond is 1.107 Å) has a small effect on the ONIOM transition properties since the trend is the same for the *high* and *low* level methods, especially in the range ±0.1 Å. Figure 7 reports the transition energy shift and the variation of the oscillator strength for the ketene. CAM-B3LYP is used in the *low* level, as this functional showed the best performance for this system. $f^{high}_{model}$ and $f^{low}_{model}$ obtained with $r_{LA}$ distorted from $r^0_{LA}$ ($r^0_{LA}$ = 1.102 Å) decrease considerably because of the mixing with other excited states of the same symmetry and close in energy. The geometry at $r^0_{LA}$ is very close to the $C_{2v}$ symmetry (the other C−H bond is 1.085 Å). The link atom slightly distorts the more symmetric geometry (the point group of the *model* system is $C_s$ as in the *real* system) in order to represent the presence of the substituent group. However, further distortion produces the mixing of various states that is responsible for the underestimated values of $f^{high}_{model}$ and $f^{low}_{model}$ away form $r^0_{LA}$. The values of $f^{high}_{model}$ and $f^{low}_{model}$ at the optimized $C_{2v}$ geometry are 1.0038 and 0.8679, respectively, which are indeed very close to those reported in Table 6. Nonetheless, the dependence on $r_{LA}$ is small in the ONIOM calculation.

## 5. Discussion and Conclusion

In this paper, we report calculations of oscillator strength with the ONIOM hybrid method. We consider four electronic excitations to valence states and compare the ONIOM results with the *target* calculation (EOM-CCSD on the entire system). We test various *low* levels of theory and check the dependence of the ONIOM results on the choice of the link atom bond length. We only define *model* systems that follow the ONIOM guidelines for partitioning.[15]

The conclusions in this paper closely follow our findings for the transition energy.[16,17] We note that it is important to verify that the same states from the subcalculations are integrated. The *low* level method that gives the best balance between accuracy and computational savings is again TD-DFT, although it is important to choose the right functional for a particular transition as shown by the ketene case. CIS and TDHF often provide poor results that make the improvement over $f^{high}_{model}$ not impressive, although an improvement over $f^{low}_{real}$ is obtained almost always. EOM-CCSD as a *low* level with a smaller basis set than the *target* usually provides the results closest to the *target*, but the $f^{low}_{real}$ (and $\Delta E^{low}_{real}$) calculation can be very demanding. Additionally, we find a case (the aldehyde) where the basis set for EOM-CCSD in

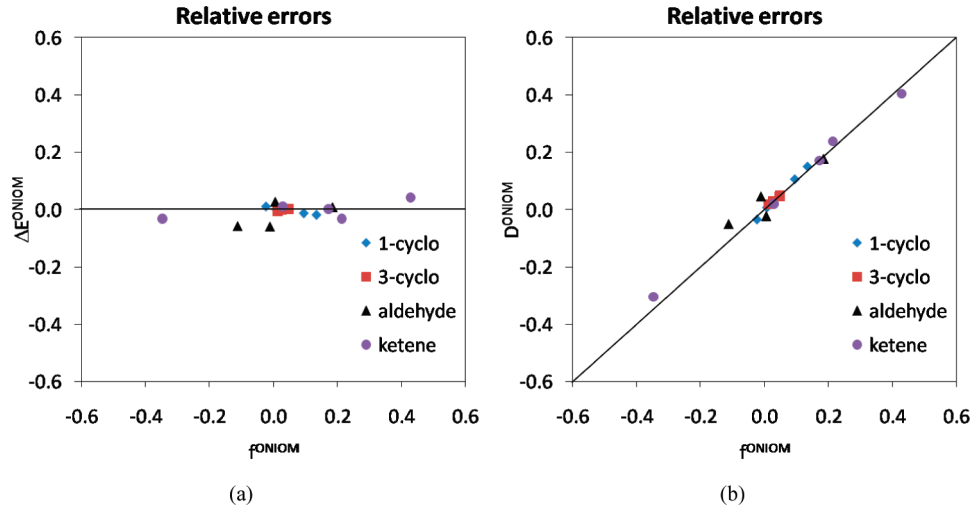



***Figure 8.*** Relative errors of the transition energy (a) and dipole strength (b) compared to the relative errors of the oscillator strength. The ketene set includes the CAM-B3LYP results.

the *low* level calculation is not adequate and provides poor results. Therefore, the choice of the basis set is very important.

Relative timings are not reported in this paper, and we refer to ref 16 for a more detailed discussion. We only point out that the bottleneck calculation is $\Delta E^{\text{high}}_{\text{model}}$ for the systems presented here when a method such as TDDFT is used in the *low* level. Since a method like EOM-CCSD scales as $O(N^6)$, where $N$ is the number of basis functions, it is evident that a hybrid method such as ONIOM can greatly reduce the computational effort by performing the expensive calculation only on the *model* system.

We confirmed that the definition of the link atom bond length (that follows the ground state formula) is appropriate also for this transition property. In fact, the dependence of $f^{\text{ONIOM}}$ (as well as of $\Delta E^{\text{ONIOM}}$) on the exact value of $r_{\text{LA}}$ is small due either to a cancellation of errors between the *model* systems subcalculations or to a small dependence of the individual subcalculations.

Although the guidelines reported in refs 16 and 17 for the transition energy also hold for the oscillator strength, the latter is a much more sensitive quantity. This is shown in Figure 8, which reports the relative errors of $\Delta E^{\text{ONIOM}}$ and $D^{\text{ONIOM}}$ as a function of the relative errors of $f^{\text{ONIOM}}$ for all the *low* level methods. These graphs show that the relative errors for the oscillator strength are larger than those of the transition energy (the graphs also include the *low* level methods that do not perform very well as discussed in the previous section). Additionally, there is a direct relationship between the dipole and oscillator strengths errors. This is not surprising, as $D^{\text{ONIOM}}$ is a quadratic quantity that depends on the independent extrapolation of the various components of the transition dipoles, eqs 4 and 5.

Our preliminary results therefore show that, although oscillator strengths are more sensitive than transition energies, they can be accurately computed with ONIOM provided that a sensible choice of the partitioning and of the *low* level method is made.

**Supporting Information Available:** Geometries of the aldehyde and the ketene and of their respective *model* systems (for the geometries of the cyclopropenes, we refer to the supporting material of ref 16). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(2) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(3) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718–730.

(4) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(5) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461*, 1–21.

(6) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(7) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.

(8) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419–1432.

(9) Vreven, T.; Morokuma, K. In *Annual reports in computational chemistry*; Elsevier: Amsterdam, 2006; Vol 2, Chapter 3, pp 35−51.

(10) Morokuma, K.; Musaev, D. G.; Vreven, T.; Basch, H.; Torrent, M.; Khoroshun, D. V. *IBM J. Res. Dev.* **2001**, *45*, 367–395.

(11) Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760–769.

(12) Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.

(13) Bearpark, M. J.; Ogliaro, F.; Vreven, T.; Boggio-Pasqua, M.; Frisch, M. J.; Larkin, S. M.; Morrison, M.; Robb, M. A. *J. Photochem. Photobiol., A* **2007**, *190*, 207–227.

(14) Hall, K. F.; Vreven, T.; Frisch, M. J.; Bearpark, M. J. *J. Mol. Biol.* **2008**, *383*, 106–121.

(15) Clemente, F. R.; Vreven, T.; Frisch, M. J. In *Quantum Biochemistry*; Wiley-VCH: Weinheim, Germany, 2010; Vol. 1, Chapter 2, pp 61−83.

(16) Caricato, M.; Vreven, T.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Phys.* **2009**, *131*, 134105.

(17) Caricato, M.; Vreven, T.; Trucks, G. W.; Frisch, M. J. *J. Chem. Phys.* **2010**, *133*, 054104.

(18) Sekino, H.; Bartlett, R. J. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1984**, *18*, 255–265.

(19) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57–62.

(20) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.

(21) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

(22) Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 9257–9269.

(23) Monkhorst, H. J. *Int. J. Quantum Chem.* **1977**, *Y11*, 421–432.

(24) Koch, H.; Jorgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333–3344.

(25) Derat, E.; Bouquant, J.; Humbel, S. *THEOCHEM* **2003**, *632*, 61–69.

(26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(27) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(28) Stephens, P. J.; Devlin, F. J.; Ashvar, C. S.; Chabalowski, C. F.; Frisch, M. J. *Faraday Discuss.* **1994**, *99*, 103–119.

(29) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(30) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

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

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

(33) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(34) Tawada, T.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.

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

(36) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.

CT1006289

JCTC Journal of Chemical Theory and Computation

# A Replica Exchange Molecular Dynamics Simulation of a Single Polyethylene Chain: Temperature Dependence of Structural Properties and Chain Conformational Study at the Equilibrium Melting Temperature

Ting Li,[†] Xiaozhen Yang,[‡] and Erik Nies*[,†,§]

*Polymer Research Division, Department of Chemistry, The Leuven Mathematical Modeling and Computational Science Centre (LMCC) and the Leuven Materials Research Centre (LMRC), Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium, State Key Laboratory of Polymer Physics & Chemistry, Center for Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Zhongguancun, Beijing 100080, Peoples' Republic of China, Laboratory of Polymer Technology, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands*

Received September 10, 2010

**Abstract:** The conformational properties of a finite length polyethylene chain were explored over a wide range of temperatures using a replica exchange molecular dynamics simulation providing high quality simulation data representative for the equilibrium behavior of the chain molecule. The radial distribution function (RDF) and the structure factor $S(q)$ of the chain as a function of temperature are analyzed in detail. The different characteristic peaks in the RDF and $S(q)$ were assigned to specific distances in the chain and structural changes occurring with the temperature. In $S(q)$, a peak characteristic for the order in the solid state was found and used to determine the *equilibrium melting temperature*. A detailed scaling analysis of the structure factor covering the full $q$ range was performed according to the work of Hammouda. In the $\Theta$ region, a quantitative analysis of the full structure factor was done using the equivalent Kuhn chain, which enabled us to assign the $\Theta$ region of our chain and to demonstrate, in our particular case, the failure of the Gaussian chain approach. The chain conformational properties *at the equilibrium melting temperature* are discussed using conformational distribution functions, using the largest principal component of the radius of gyration and shape parameters as order parameters. We demonstrate that for the system studied here, the Landau free energy expression based on this conformational distribution information leads to erroneous conclusions concerning the thermodynamic transition behavior. Finally, we focus on the instantaneous conformational properties *at the equilibrium melting temperature* and give a detailed analysis of the conformational shapes using different shape parameters and a simulation snapshot. We show that the chain does not only take the lamellar rod-like and globular conformational shapes, typical of the solid and liquid states, but can also explore many other conformational states, including the toroidal conformational state. It is the first demonstration that a flexible molecule like PE can also take a toroidal conformational state, which is normally linked to stiffer chains.

## Introduction

The single chain behavior is not only elementary in understanding the properties of polymers in the condensed state but also has immediate relevance in nanoscience, nanotechnology, and experimental techniques, such as AFM, which allows for the direct study of single chains. The study of the conformational properties of a single chain by experimentation,[2−19] theory,[20−23] and simulation[24−39] has a long-standing history in fundamental polymer science. It is well established that a polymer chain in dilute solution can be present as an expanded coil, an ideal coil, or a

* Corresponding author. Tel.: +32 16 327481or +32 16 327418. Fax: +32 16 327990. E-mail: Erik.Nies@chem.kuleuven.be.
[†] Katholieke Universiteit Leuven.
[‡] Chinese Academy of Sciences.
[§] Eindhoven University of Technology.

10.1021/ct100513y 
Published on Web 12/09/2010

collapsed globule depending on the temperature, solvent quality, and pressure.[24,40,2,41−43] Moreover, the expanded coil, the ideal coil, and the globule are not the only possible conformational states, and depending on the details of the intrachain and chain−solvent interactions, other conformational states may exist.[44−47] For instance, for stereo-regular polymers, the additional possibility exists for it to occur in the crystalline state, and the single chain can attain a folded lamellar crystalline conformational state.

Simulation and theory have contributed a great deal to the understanding of the existence of different single chain conformational states and transitions between them.[24,48,49] Initially, attention was given to the coil−globule transition.[50,51] More recently, the low temperature behavior received increased attention in computer simulation studies, and e.g., liquid−solid and solid−solid transitions have been observed.[52−56] The transitions between the coil, globule, and folded chain states are affected by both the environmental variables and the intrinsic character of the polymer chain. For instance, the occurrence of the globular state upon going from the expanded coil to the chain folded crystal and the length/diameter aspect ratio of the chain folded crystal sensitively depends on the chain stiffness[44,57,58] and the chain length.[48]

Unfortunately, computer simulation studies of the dense chain conformations at low temperatures are far from easy. Different instantaneous chain folded conformational states are very likely separated by high free energy barriers, making it very difficult to realize transitions from one state to the other in the time of a single Molecular Dynamics or conventional Monte Carlo simulation run. Therefore, the simulation does not necessarily represent a proper ensemble average over all relevant chain conformational states, and as a consequence, the simulation results are not necessarily representative of the behavior of real chains observed in typical experiments in which the time of the experiment is sufficiently long and/or the number of chains probed by the experiment is sufficiently large to properly sample the different conformational states. To overcome the problem of poor sampling of the rugged free energy landscape, advanced simulation methods have been developed. In particular, expanded ensemble simulation methods, such as the replica exchange method (REM) or parallel tempering (PT),[59−67] multicanonical ensemble method (MUCA),[68−70] and four dimensional expanded ensemble algorithm,[71,72] have proven to be very efficient.

In this work, we apply, for a polyethylene chain of finite length, canonical MD simulations combined with the replica exchange method (REMD)[59,60,73] to explore a wide range of temperatures, enabling us to efficiently sample the whole relevant phase space and to provide simulation data of high quality, which are representative for the equilibrium behavior of the chain molecule at all studied temperatures. From the simulation data, we obtain the radial distribution function and the structure factor of the chain as a function of the temperature and analyze these statistical properties in detail. The structure factor is the key property obtained from scattering experiments and contains statistical information on the structure of the chain molecule, and therefore this analysis may be of use in the interpretation of such experimental data.

Subsequently, we discuss the distribution function of the largest principal component of the radius of gyration *at the equilibrium melting temperature* and discuss the applicability and the pitfalls of the use of a Landau free energy expression based on this conformational distribution information. Finally, we focus on the instantaneous conformational properties *at the equilibrium melting temperature* and give a detailed analysis of the conformational shapes using different shape parameters.

The paper is organized as follows. First the model and simulation parameters are described. Then, the structural and conformational analysis of the simulation results is presented and discussed. Finally, some conclusions are made.

## Simulation Details

In this study, a linear polyethylene chain consisting of $N = 200$ $CH_2$groups was modeled as a bead−spring chain using the united atom (UA) approximation for the $CH_2$ units and making no distinction between middle and end groups. The united atom approximation is widely used in the simulation of macromolecules.[25,52,74,75] Four types of potentials are included in our simulation, i.e., bond stretching, angle bending, torsional rotation, and van der Waals (vdW) interactions between atoms separated more than two covalent bonds along the polymer chain (i.e., including the 1−4 interaction). The potential energy expressions and the parameter values for the united atom model are taken from the DREIGING force field:[76]

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}} + U_{\text{vdw}} \tag{1}$$

in which

$$U_{\text{bond}} = \sum \frac{1}{2} K_b (l - l_0)^2,$$

$$U_{\text{angle}} = \sum \frac{1}{2} K_a (\theta - \theta_0)^2,$$

$$U_{\text{torsion}} = \sum \frac{1}{2} K_t \{1 - \cos[n(\phi - \phi_0)]\}$$

and

$$U_{\text{vdW}} = \sum \sum 4\varepsilon \left[ \left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]$$

(The summations denote that the potential energy terms are summed over all bonds, angles, torsion angles, and nonbonded pairs, respectively), and parameter values are given in Table S1 of the Supporting Information.

In the following, we use reduced units. The van der Waals diameter $\sigma$, the energy parameter $\varepsilon$, and the mass of the united atom are taken to define reduced units that are denoted by the superscript asterisk, for example, reduced temperature $T^* = k_B T/\varepsilon$, reduced density $\rho^* = \rho\sigma^3$, reduced time $t^* = t\sqrt{(\varepsilon/m)}/\sigma$, reduced distance $r^* = r/\sigma$, and reduced amplitude of the scattering vector $q^* = q\sigma$. For the particular parameter values used in this study, the following conversions are obtained: $t = 1.488 \times t^*$ with the time $t$ in picoseconds and $T = 99.921 \times T^*$ with the temperature $T$ in degrees Kelvin.

A REMD or parallel tempering molecular dynamics (PTMD) simulation with in total 77 temperatures, covering a wide temperature range from 252 to 1511 K, was executed using at each temperature a massive Nose−Hoover chain (MNHC) thermostat.[77] The reduced time step was $\Delta t^* = 1.6427 \times 10^{-3}$ (or 2.444 fs in real units). Methods to optimize the REM parameters (such as the number of replicas or the temperature distribution) have been discussed recently by Trebst et al.,[78] Katzgraber et al.,[79] Nadler and Hansmann,[80−82] and others.[83−85] In these studies, fine-tuned parameter sets and optimization strategies have been demonstrated to be useful in improving the efficiency of REM simulation. In the present work, we did not use sophisticated optimization methods. Instead, the temperatures were determined in preliminary trial REMD simulation runs, ensuring that the potential energy probability density functions (PDF) for any two adjacent temperatures have considerable overlap, such that sufficiently high swapping rates between temperatures are achieved.[86] The used set of reduced temperature values are given in Table S2 of the Supporting Information.

In the Supporting Information, we make a safe (conservative) estimate of the number of independent conformations contributing to the simulation results. Typically, we find that at each temperature more than 1000 truly independent conformations contribute to the simulation averages determined over large number of snapshots, typically $10^7$ in total. More details about the parameters, the performance, and the statistics of the REMD simulation are also discussed in the Supporting Information.

## Results and Discussion

**1. Temperature Dependence of the Radial Distribution Function (RDF).** In this work, the radial distribution function (RDF) is used to investigate the intramolecular structure of the single chain. In Figure 1a and b, the equilibrium RDF is presented at different temperatures; in Figure 1a, a 3D plot is given at all simulated temperatures.

The first and highest peak in the RDF is due to the bond length distribution and appears for all temperatures at $r^* \cong 0.422$. The maximum of the second peak, situated at $r^* \cong 0.690$ at the lowest temperature and slightly shifted to $r^* \cong 0.706$ at higher temperatures, is due to the distance between 1−3 atoms (two atoms bonding to a common atom) and is close to the value $r^*_{13} \cong 0.689$, calculated from the Dreiding force field parameters. The shift of this peak indicates that the equilibrium bond angle in the polymer chain increases slightly as the temperature increases. The maximum of the third peak located at $r^* \cong 1.06$ at $T^* \cong 2.52$ shifts to a smaller value $r^* \cong 1.04$ at the highest temperature $T^* = 15.12$. At the high temperatures, this peak also gets a shoulder at shorter distances. The reduced distances between 1−4 atoms (two atoms separated by three consecutive bonds along the chain) in the *trans* and the *gauche* conformations calculated from the force field are ca. 1.063 and 0.808, respectively. The *trans* state is energetically more favorable than the *gauche* state and, hence, more probable at lower temperatures. At higher temperatures, the *gauche* state becomes more populated and gives rise to the shoulder at a slightly shorter distance in the RDF. Next to the first three peaks, in Figure 1b, there are more peaks appearing at larger distances

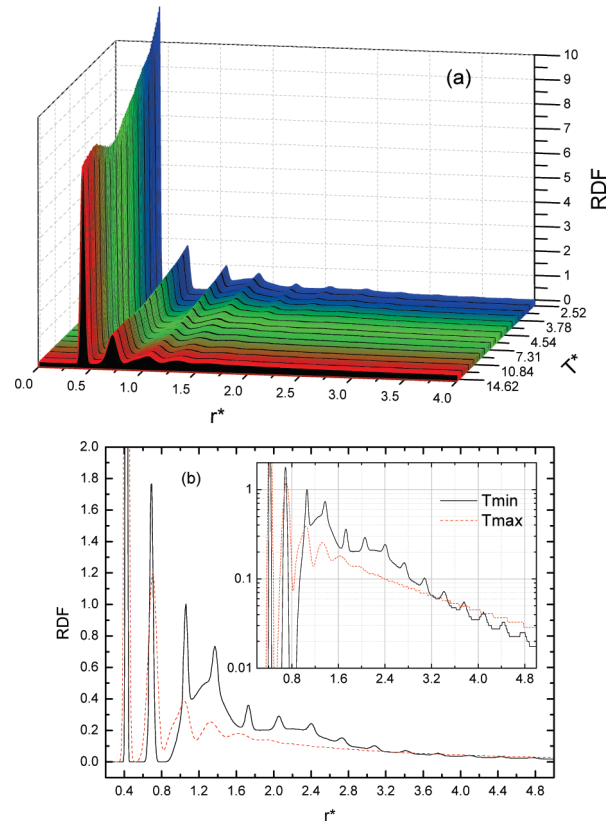



***Figure 1.*** Radial distribution function of the PE chain. (a) Plot of RDF versus $r^*$ and reduced $T^*$. (b) Detail of the RDF vs $r^*$ at the lowest temperature ($T^*_{min} = 2.52$) and the highest temperature ($T^*_{max} = 15.12$).

which are well developed at the lowest temperature $T^* = 2.52$. The maxima of these regularly spaced peaks correspond to distances between two atoms separated by four or more consecutive bonds with all dihedral angles in the *trans* state. At low temperatures, the chain is in the folded lamellar state,[86] and the sharpening of these regularly spaced peaks at lower temperatures indicates the growing and perfecting of the *all-trans* stems in the chain folded structure. In an *all-trans* stem, all of the torsion angles formed by at least four consecutive chain units fall in the angular ranges $[-\pi,-5\pi/6)$ or $(5\pi/6,\pi]$. The smearing out of the peaks with increasing temperature indicates the shortening and disordering of the *all-trans* stems accompanying the melting of the lamellar crystal, as discussed previously.[86] In Figure 1b, the RDFs at the lowest ($T^* = 2.52$) and the highest ($T^* = 15.12$) temperatures are compared. Here, it is noticeable that the sharp *all-trans* peaks are superimposed on a broader underlying peak.

To better understand the origin of the broadened peak in the RDF in Figure 1b, we calculate a modified RDF in which we systematically exclude more and more pair distances from the calculation. In Figure 2a and b, we show the RDFs calculated at $T^* = 2.52$ and $T^* = 15.12$, respectively, with an increasing number of excluded pair distances. For example, in Figure 2, from bottom to top, curve 1 is the full RDF at the corresponding temperature, curve 2 is calculated with all pairs $[i,i+1]$ excluded, curve 2 is calculated with all pairs $[i,i+1]$ and $[i,i+2]$ excluded, and so on, until in curve 10, the pairs $[i,i+1]$, $[i,i+2]$, ..., $[i,i+8]$, $[i,i+9]$ are excluded. In this way, we can hide the contributions of

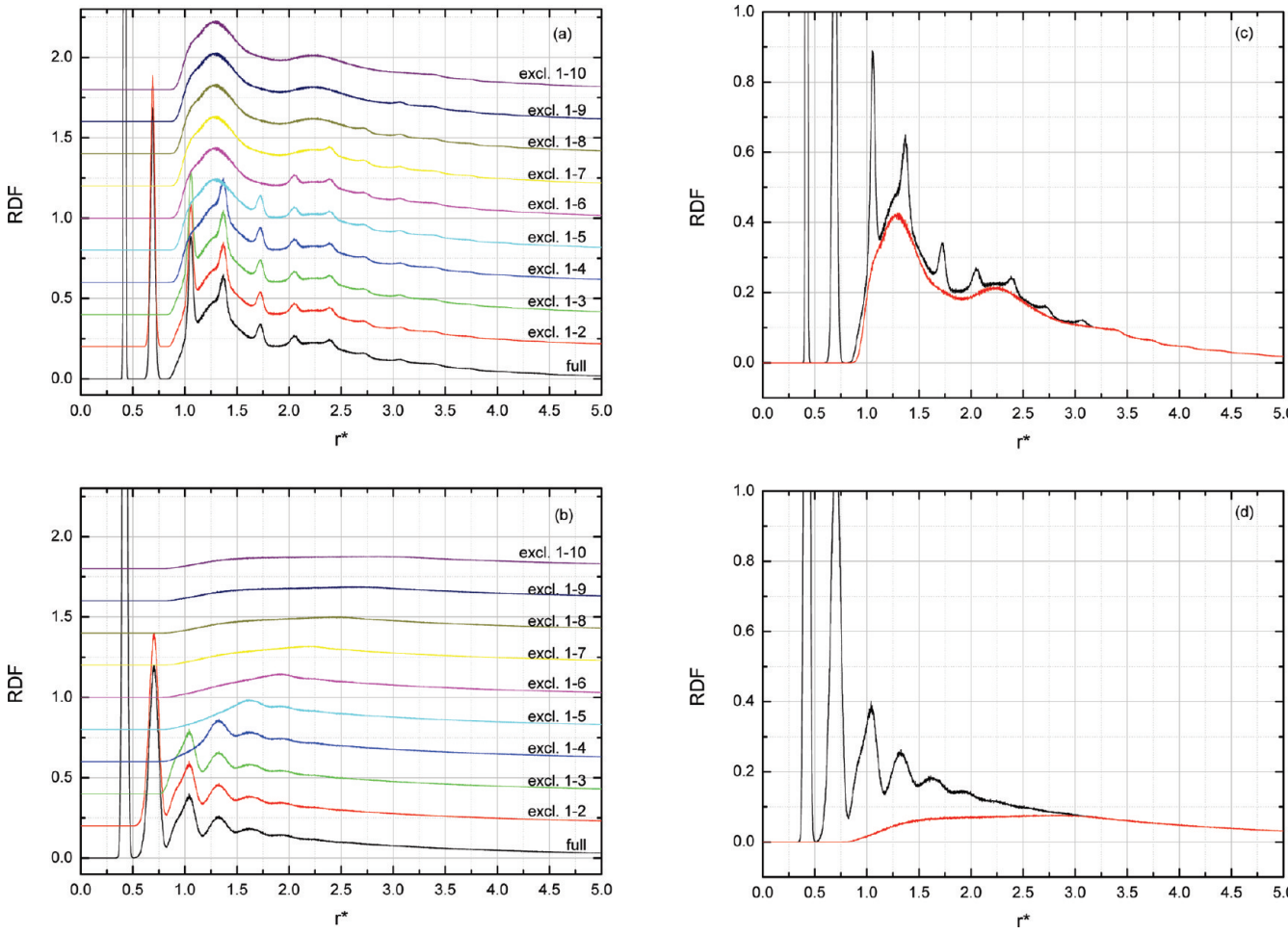
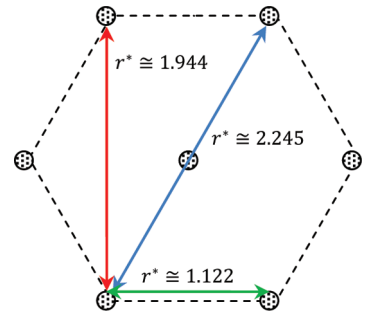



***Figure 2.*** Full RDFs and partial RDFs including only nonbonded pairs that are at least separated by *n* number of consecutive bonds along the chain with *n* from 1 to 9 at the lowest temperature ($T^*_{min} = 2.52$) and the highest temperature ($T^*_{max} = 15.12$). (a) A stack plot showing the full RDF and partial RDFs with all of the exclusions at $T^*_{min} = 2.52$. (b) A stack plot showing the full RDF and partial RDFs with all of the exclusions at $T^*_{max} = 15.12$. (c) Comparison of the full RDF and the partial RDF with the maximal exclusion at $T^*_{min} = 2.52$. (d) Comparison of the full RDF and the partial RDF with the maximal exclusion at $T^*_{max} = 15.12$.

specific pair distances to the total RDF. As expected with an increasing number of excluded pair distances, the sharp peaks in the full RDF are removed, and we are left with the RDF due to pair distances farther apart along the contour of the chain. In Figure 2c and d, the full RDF and the RDF excluding all pairs up to $[i,i+9]$ are presented in one plot for $T^* = 2.52$ and $T^* = 15.12$. The main finding here is that the nonbonded RDFs are very different for the two temperatures. At low temperatures, the nonbonded RDF clearly shows two peaks, whereas at the high temperature RDF, only one very broad smeared peak remains. At $T^* = 2.52$, the two rounded peaks at $r^* \cong 1.297$ and $r^* \cong 2.235$ are dominated by $CH_2$ units far apart along the chain contour but that are nearest neighboring and next-nearest neighboring $CH_2$ units in the hexagonal lamellar folded crystal. Scheme 1 gives a schematic of the cross-section perpendicular to the stems in the chain folded crystal, illustrating the hexagonal packing of the $CH_2$ units (represented by the shaded dots). The reduced distances between the nearest ($r^* \cong 1.122$) and next nearest neighboring ($r^* \cong 1.944$ and $r^* \cong 2.245$) $CH_2$ units presented in Scheme 1 are calculated from the force field parameters assuming perfect hexagonal packing of

***Scheme 1.*** Schematic Cross-Section Perpendicular to the Stems in the Chain Folded Crystal, Illustrating the Hexagonal Packing of the $CH_2$ Units in the *All-trans* Stems[a]




[a] The reduced distances between the nearest ($r^* \cong 1.122$) and next nearest neighboring ($r^* \cong 1.944$ and $r^* \cong 2.245$) $CH_2$ units are calculated from the force field parameters assuming perfect hexagonal packing of stems.

stems. In Figure 2c, a clear shoulder around $r^* \cong 1.1$ can be seen in the first underlying peak with a maximum at $r^* \cong 1.297$. The shoulder stems from $CH_2$ units at the vdW nearest neighbor distance expected from the pair potential.

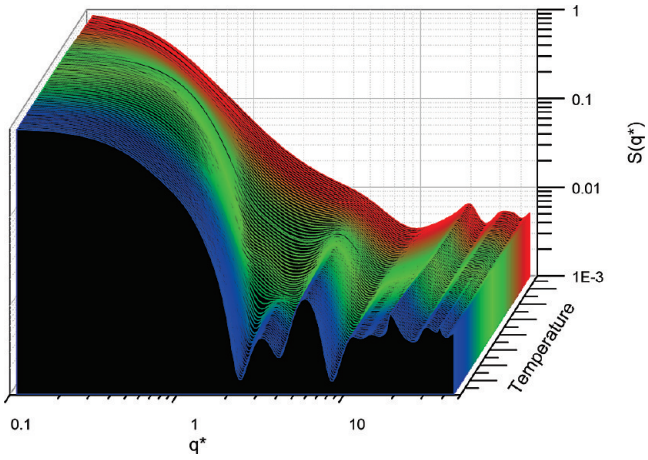



***Figure 3.*** The static structure factor $S(q^*)$ as a function of reduced $q^*$ at the 77 simulation temperatures ranging from $T^*_{min} = 2.52$ to $T^*_{max} = 15.12$.

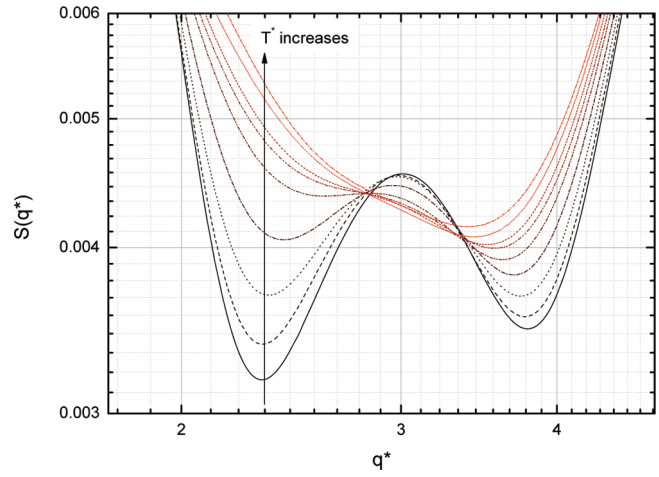



***Figure 4.*** $S(q^*)$ around $q^* \cong 3.189$ at temperatures $T^* = 4.083$, 4.133, 4.183, 4.234, 4.284, 4.335, 4.385, 4.435, and 4.486 (corresponding to the curves from bottom to top at $q^* = 2.3$).

However, the maximum of the peak is at $r^* \cong 1.297$, indicating that other distances of nonbonded pairs also contribute to this peak. The chain connectivity leads in the crystalline state to a tilting of stems, $CH_2$ units in folds, etc. and makes it so that not all nearest neighbors are located at the most optimal vdW distance. The peak observed at $r^* \cong 2.235$ is attributed to next nearest neighbors and contains $CH_2$ units present in the second neighbor shell ($r^* \cong 1.944$) and is attributed to first shell $CH_2$ units in diagonal positions compared to the central $CH_2$ unit ($r^* \cong 2.245$).

In Figure 2d, at the highest temperatures $T^* = 15.12$, the chain is in the expanded coil state and the RDF shows a very broad peak which is characteristic for the gaseous-like density and distribution of the nonbonded pairs.

**2. Temperature Dependence of the Static Structure Factor $S(q^*)$.** The single-chain static structure factor for an isotropic system is defined as

$$S(q^*) = \frac{1}{N^2}\left\langle \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\sin(q^* R^*_{ij})}{q^* R^*_{ij}} \right\rangle \qquad (2)$$

where the brackets represent the time average in the MD simulation. In Figure 3, $S(q^*)$ patterns at all temperatures used in the REMD simulation are presented in a 3D plot. We will analyze and discuss the relevant regimes of $S(q^*)$ in detail and relate them to the structural changes with the temperature occurring in the chain.

The two shortest length scales in real space are the distances between adjacent $CH_2$ units (the covalent bond length $l^*_0 = 0.422$) and the distance between two units connected to a common $CH_2$ unit (the distance related to the bond angle $r^*_0 \cong 0.689$), corresponding in $q^*$ space to $q^* = 2\pi/l^*_0 \cong 14.9$ and $q^* = 2\pi/r^*_0 \cong 9.12$. We have ascertained that the detailed shape of $S(q^*)$ in the large $q^*$ regime, viz, $q^* > 2\pi/0.689 \cong 9.12$, is primarily determined by the bond length and bond angle distances, which vary little with the temperature, as can be confirmed in Figure 3.

In the range $2\pi/\sqrt{\langle R^{*2}_g\rangle} < q^* < 9.12$, both local and larger scale structure information and scaling behavior for different chain states are contained. As the temperature decreases, in this $q^*$ range, two peaks emerge (see Figure 3) which, at the lowest temperature $T^* = 2.52$, are located at $q^* \cong 3.189$ and $q^* \cong 5.816$. The corresponding distances in real space using Bragg's law are $r^* \cong 1.970$ and $r^* \cong 1.080$, respectively. These distances are close to the values calculated on the basis of the force field parameters for the next nearest and the nearest neighboring *all-trans* stems (see Scheme 1). The peak at $q^* \cong 5.816$ is related to the nearest nonbonded pairs and is present in the crystalline as well as the globular state and is characteristic of the high segmental density in these two states. In the expanded coil state, this peak disappears as the distribution of the chain segments is more typical of a gaseous state. The peak at $q^* \cong 3.189$ corresponding to next nearest neighboring distances is only present in the crystalline state and is representative of the ordered state. The appearance of this peak can be used to determine the equilibrium melting temperature. The *equilibrium melting temperature* and the *equilibrium lamellar thickness* of the chain folded lamellar crystals have been determined previously for a PE chain with the same chain length, using the same force field. The equilibrium melting temperature was estimated from different properties: in particular, the peak position of the total heat capacity as well as the heat capacities due to van der Waals and torsional interactions were related to the solid−liquid equilibrium; the changes of the radius of gyration and of the orientational order parameters of the *all-trans* stems with temperature were used to estimate the equilibrium melting temperature.[86] The estimates of the equilibrium melting temperature from these different properties all gave the same value for the equilibrium melting temperature (in reduced units, $T^0_m{}^* = 4.234$).

In Figure 4, we show a detail of $S(q^*)$ around $q^* \cong 3.189$ (peak position for $T^*_{min} = 2.52$) in the vicinity of the equilibrium melting temperature. We see that a peak exists at lower temperatures with its maximum at $q^* \approx 3$ slightly shifting to lower $q^*$ as the temperature increases. The $S(q^*)$ peak height at $q^* \approx 3$ increases rapidly in the vicinity of $T^0_m{}^*$, indicating that at this temperature the second nearest neighboring order, characteristic of the crystalline state, rapidly changes at the equilibrium melting temperature. Hence, $S(q^*)$ can also be used to determine the equilibrium

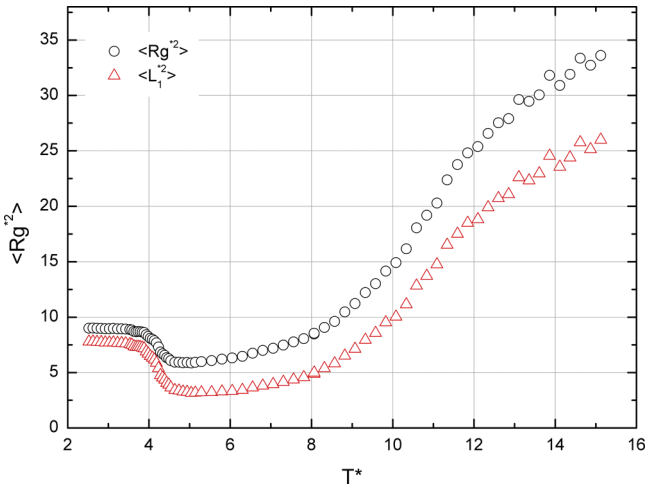



***Figure 5.*** Mean squared radius of gyration $\langle Rg^{*2}\rangle$ (circles) and largest principal component of the radius of gyration $\langle L_1^{*2}\rangle$ (triangles).

melting temperature, leading to an estimate for $T_m^0{}^*$ in agreement with the estimate for $T_m^0{}^*$ obtained from other independent methods.[86]

**3. Scaling Behavior of *S*(*q**).** In Figure 5, the temperature dependence of $\langle Rg^{*2}\rangle$ and the largest principal component of the radius of gyration $\langle L_1^{*2}\rangle$ calculated from the simulation data is shown; the details have been discussed elsewhere.[86]

In Figure 6a−d, the $S(q^*)$ curves are presented at five selected temperatures, namely, $T^* = 2.52$, 5.04, 10.84, 12.60, and 15.12.

At the lowest temperature $T^* = 2.52$ and the highest temperature $T^* = 15.12$ used in the REMD simulation, the chain is in the folded chain and the expanded coil state, respectively. In Figure 5, at $T^* = 5.04$, the chain has the smallest $\langle Rg^{*2}\rangle$, and at $T^* = 10.84$, $\langle Rg^{*2}\rangle$ shows an inflection point, which is an indication of the transition between the globule and coil states, although it does not define the exact Θ state.[87] Finally, $T^* = 12.60$ is selected because there is a linear part in the $S(q^*)$ where it has a slope of −2, characteristic of ideal coil behavior.

In the Porod regime, $q^* > 2\pi/\sqrt{\langle Rg^{*2}\rangle}$, the structure factor may show scaling behavior, $S(q^*) \sim (q^*)^{-\alpha}$ with α the Porod exponent. The value of α can be linked to the shape or the fractal character of a polymer chain. In a log−log plot of $S(q^*)$ versus $q^*$, the scaling behavior becomes clear as a linear region of $S(q^*)$ with a slope of −α.

In a recent study, Hammouda presented a new Porod-like analysis to interpret the structure factor obtained in scattering experiments and applied it to small-angle neutron scattering data of polymers in a solution forming micelles of various shapes, including the sphere, the cylinder, and the lamella. His analysis shows that for a cylindrical micelle three scaling regimes exist with $\alpha = 0$, 1, and 4. It is known that for three-dimensional objects with smooth surfaces, the Porod exponent $\alpha = 4$, whereas for a one-dimensional rigid rod, it should be 1.[1,88] At the lowest temperature ($T^* = 2.52$), the chain folds into a cylinder-like shape with rough surfaces and $S(q^*)$ is flat ($\alpha = 0$) at very small $q^*$; at intermediate $q^*$ values ($0.1 < q^* < 0.2$), $S(q^*) \sim (q^*)^{-1}$ can be observed. As $q^*$ further increases, an $\alpha = 4$ scaling is reached. Following Hammouda, this behavior indicates that at this temperature the chain is present as a cylinder with finite thickness. When $T^* = 5.04$, $\langle Rg^{*2}\rangle$ has the smallest value; the chain is in a disordered but compact globular state. The $S(q^*)$ (Figure 5a) shows an extended flat region at low $q^*$ and then quickly drops with exponent $\alpha = 4$ and can be seen to reflect that the chain is spherical with a relative smooth surface formed.[1] At the highest simulation temperature $T^* = 15.12$, see Figure 5b, the $S(q^*)$ data follow a scaling law with $\alpha \approx 5/3$ approximately, which indicates that the chain behaves like an excluded volume coiled chain.[88] At $T^* = 12.60$, see Figure 5b, the $S(q^*)$ data follow scaling behavior $S(q^*) \sim q^{*-2}$ ($\alpha = 2$), which indicates the chain behaves like an ideal coil; i.e., the chain is in the Θ state.[88]

**4. *S*(*q**) in the Θ Region.** In Figure 7, we investigate the Θ behavior in some more detail and compare the structure factor $S(q^*)$ at five temperatures from the simulation results with the structure factor calculated assuming that the chain behaves like an ideal Kuhn chain.

The structure factor of the Kuhn chain is given by

$$S(x^*) = \frac{1 + (N_K - 1)(1 - x^{*2}) - 2x^*(1 + x^* + x^{*N_K})}{N_K^2(1 - x^*)^2}$$

with $x^* = (\sin q^* l_K^*)/(q^* l_K^*)$, $N_K$ being the number of Kuhn segments, and $l_K$ being the Kuhn length.

The number of Kuhn segments, and the Kuhn length, are calculated from the end-to-end distance of the chain $\langle R^2\rangle$ and the contour length $L = (N - 1)\cdot\langle l_b\rangle\cdot\sin(\langle\theta\rangle/2)$, with $\langle l_b\rangle$ the effective bond length and $\langle\theta\rangle$ the effective bond angle also obtained in the simulation. Figure 7a shows $S(q^*)\cdot q^{*2}$versus $q$ (Kratky plot), and the best agreement between the simulation and the Kuhn chain results is obtained for the temperatures $T^* = 12.60$ and $T^* = 12.85$. For higher and lower temperatures, the simulation data and the Kuhn result deviate significantly: the chain behaves at higher temperatures as an excluded volume chain and at lower temperatures behaves as a globule. Only at the temperatures $T^* = 12.60$ and $T^* = 12.85$ can the chain be treated as an ideal Kuhn chain of finite size: the temperature region $T^* = 12.60-12.85$ provides a first estimate for the Θ region of our chain.

In Figure 7b, $S(q^*)$ from the REMD simulation at $T^* = 12.60$, $S(q^*)$ calculated from the Debye expression for the Gaussian chain, and $S(q^*)$ of the equivalent freely jointed Kuhn chain are compared. For the Gaussian chain, the Debye result is

$$S(Q^*) = \frac{2(e^{-Q^*} - 1 + Q^*)}{Q^{*2}}$$

with

$$Q^* = \frac{q^{*2} N l_0^{*2}}{6} = q^{*2}\langle R_g^{*2}\rangle$$

Clearly for the temperatures in the Θ region, ideal chain behavior is reached with the appropriate scaling behavior $S(q^*) \sim q^{*-2}$ (denoted by the dotted straight line in Figure

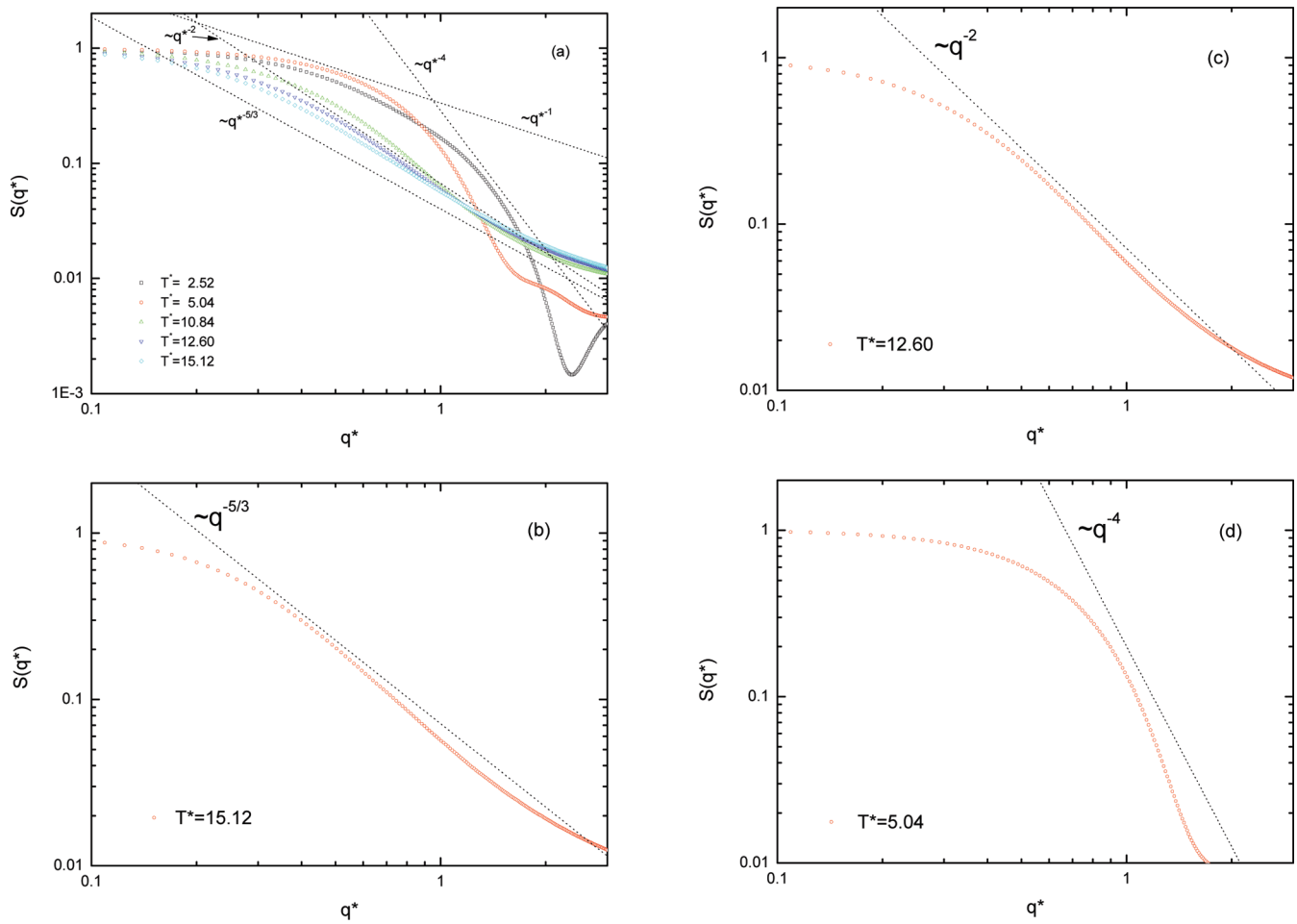



***Figure 6.*** (a) Structure factor $S(q^*)$ (shown in symbols) at five different temperatures $T^* = 2.52$ (squares), 5.04 (circles), 10.84 (upward triangles), 12.60 (downward triangles), and 15.12 (diamonds). Relevant scaling laws (dotted straight lines) are also presented. (b) Plot of $S(q^*)$ at $T^* = 15.12$. (c) Plot of $S(q^*)$ at $T^* = 12.60$ and (d) plot of $S(q^*)$ at $T^* = 5.04$.

7b). In the small $q^*$ range ($q^* < 0.5$), the simulated $S(q^*)$ is in good agreement with the calculated $S(q^*)$ based on the two models. However, in the larger $q^*$ region ($0.5 < q^* \leq 2$), the simulated $S(q^*)$ starts deviating from $S(q^*)$ for the Gaussian chain, whereas the $S(q^*)$ based on the Kuhn chain is still in quantitative agreement with the simulation data. This demonstrates that the ideal behavior of a real chain of finite length considered here is better described by the finite freely jointed chain than by the Gaussian chain model.

In experimental practice, Kratky plots are routinely used to determine the Θ state in the following way: Ideal chain behavior is assumed when in the Kratky plot $S(q^*)\cdot q^{*2}$ versus $q^*$ becomes independent of $q^*$ at intermediate $q^*$. A detailed Monte Carlo simulation study of the single chain structure factor at all length scales has been presented for sufficiently long chains under good and Θ solvents conditions.[89] These authors discussed in detail the scaling behavior from a (modified) Kratky plot under good and Θ-solvent conditions. It was clearly shown that for sufficiently long chains the (modified) Kratky plot indeed displays, in agreement with experimental practice, a plateau at intermediate $q^*$ indicative of the chain scaling behavior. The chain length discussed in this work is certainly not long enough, and therefore the standard method of analysis does not apply. In our simulation results, we see that in the Θ region the Kratky plot does not show a $q^*$ range where $S(q^*)\cdot q^{*2}$ is independent of $q^*$. The reason is the following: $S(q^*)\cdot q^{*2}$ versus $q^*$ only has a horizontal for finite length Gaussian chains or for sufficiently long ideal chains with constant bond lengths. The number of Kuhn segments in our chain is rather small, $N_K = 28-30$, and therefore the necessary condition of sufficiently long chains is not met and the Kratky plot of the ideal Kuhn chain does not have a horizontal line segment. Hence, in order to ascertain whether the chain behaves ideally, it is not sufficient to search for a straight line piece; one must also take into account the long chain condition or, for smaller chains, consider the full structure factor in the range $0 \leq q^* < l_K$. In fact, when we use only the criterion of a horizontal line piece in the Kratky representation for our simulation data, we would conclude that at $T^* = 13.86$ the Θ state is reached, a considerably higher temperature than the Θ temperature derived from scaling and the Kuhn analysis. Therefore, in the interpretation of simulation as well as experimental data, one must take care that all assumptions are obeyed in order to correctly establish ideal chain behavior.[87]

**5. Landau Free Energy at $T_m^{0*}$: Potential Energy and Radius of Gyration Distribution Functions.** It has been reported that a chain molecule can have many different conformational states, such as coiled, globular, rod-like, and

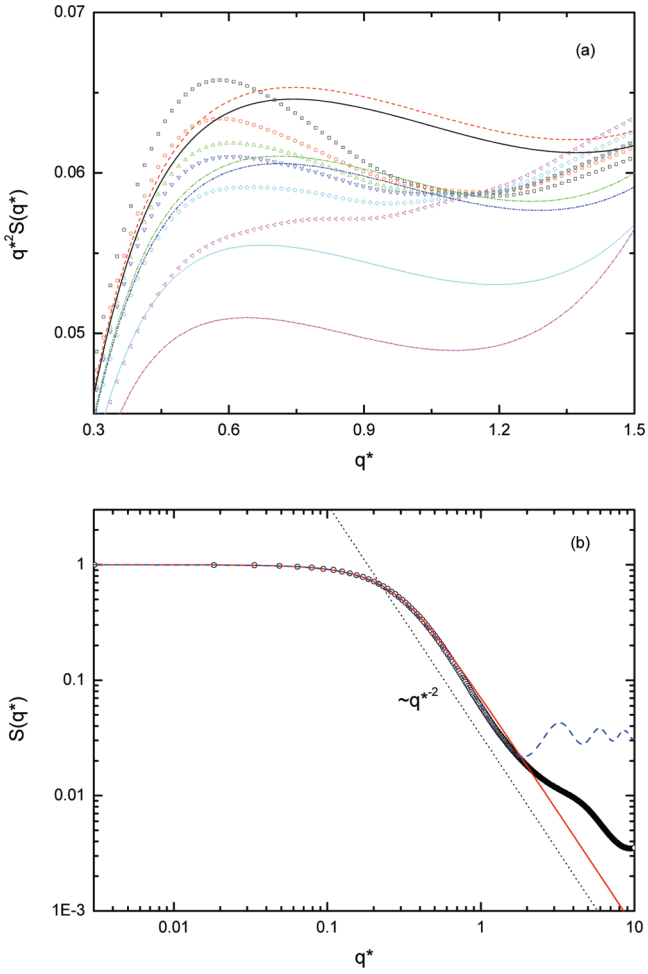



***Figure 7.*** (a) Krakty plots for $T^* = 12.10$ (squares), 12.35 (circles), 12.60 (upward triangles), 12.85 (downward triangles), 13.10 (diamonds), 13.86 (leftward triangles), together with the calculated lines based on the Kuhn model for $T^* = 12.10$ (solid), 12.35 (dash), 12.60 (dot), 12.85 (dash dot), 13.10 (dash dot dot) 13.86 (short dash). $T^* = 12.60$ and 12.85 show the best fitting to the calculated curve (in a sense of nonlinear least-squares fitting). (b) Comparison of the $S(q^*)$ of the REMD simulation at $T^* = 12.60$ (circles) and calculated $S(q^*)$ based on the Debye formula for a Gaussian chain (solid line) and the Kuhn equivalent chain (dashed line).

toriodal states, depending on, e.g., the chain stiffness.[90−100] From computer simulations of a single PE chain, it is also well established that the chain can exist in the coiled, globular, and chain folded states, but whether other conformational shapes are possible and present is not clear. Here, we study the conformational shapes of the PE chain at its equilibrium melting temperature $T_m^{0*}$ in more detail.

Figure 8 shows the probability density distribution of the potential energy, a unimodal and Gaussian-like curve.[86] However, when we study the relationship between the energy and the chain shapes, interesting results are found. In Figure 8b, the correlation between the potential energy and the largest principal component of the mean squared radius of gyration $L_1^{*2}$ at $T_m^{0*} = 4.23$ is presented in a scatter plot. Each point in the scatter plot represents one chain configuration found in the equilibrium trajectory, which contains $5 \times 10^5$ configurations in total.

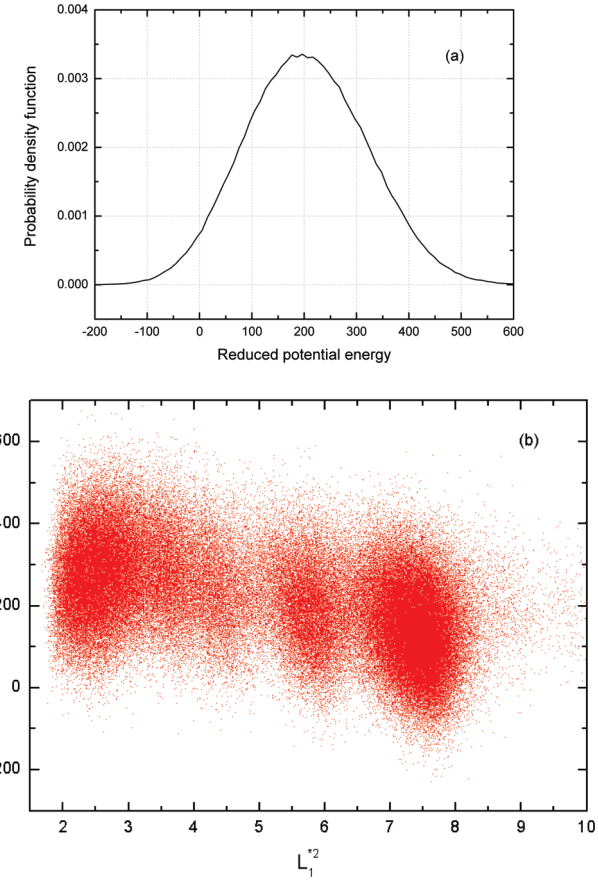



***Figure 8.*** (a) Probability density distribution of the reduced potential energy and (b) correlation between the reduced potential energies $U^*$ and the largest principal component of radius of gyration $L_1^{*2}$ of $5 \times 10^5$ chain configurations at the crystallization temperature $T_m^{0*} = 4.23$. (b) It is clearly shown that the population of conformations is unevenly and broadly distributed along $L_1^{*2}$, implying that the chain configurations can be divided into several groups on the basis of a structural parameter such as $L_1^{*2}$.

The probability density function (PDF) of $L_1^{*2}$, denoted by $P(L_1^{*2})$, is presented in Figure 9a. The multimodal feature in $P(L_1^{*2})$ has also been observed in the radius of gyration by others.[91] However, the PDF of the potential energy $P(U^*)$ is unimodal, as shown in Figure 8a. To understand how the configurations with different structural parameters contribute to the thermodynamic quantity $U^*$, the chain configurations around each peak found in $P(L_1^{*2})$ are grouped by their $L_1^{*2}$ values. Therefore, five groups are defined corresponding to the five peaks found in $P(L_1^{*2})$ (the name and boundary for each group are labeled in Figure 9a). The chain configurations in each group are extracted, and the PDF of the potential energy for the chain configurations in each group is calculated. The PDFs for the five configuration groups ($P_i(U^*)$, $i = 1-5$, with $i$ the peak index; thin lines) are presented in Figure 9b. Clearly, the energy distributions $P_i(U^*)$ for the different groups are not distinct and show considerable overlap. Peak 1 in Figure 9a, which is located at the largest $L_1^{*2}$, gives $P_1(U^*)$ at the lowest energy in Figure 9b. Peak 2, which has the smallest $L_1^{*2}$, possesses a distribution at higher energies. The energy distributions of peaks 3 and 4 are in between the distributions of peaks 1

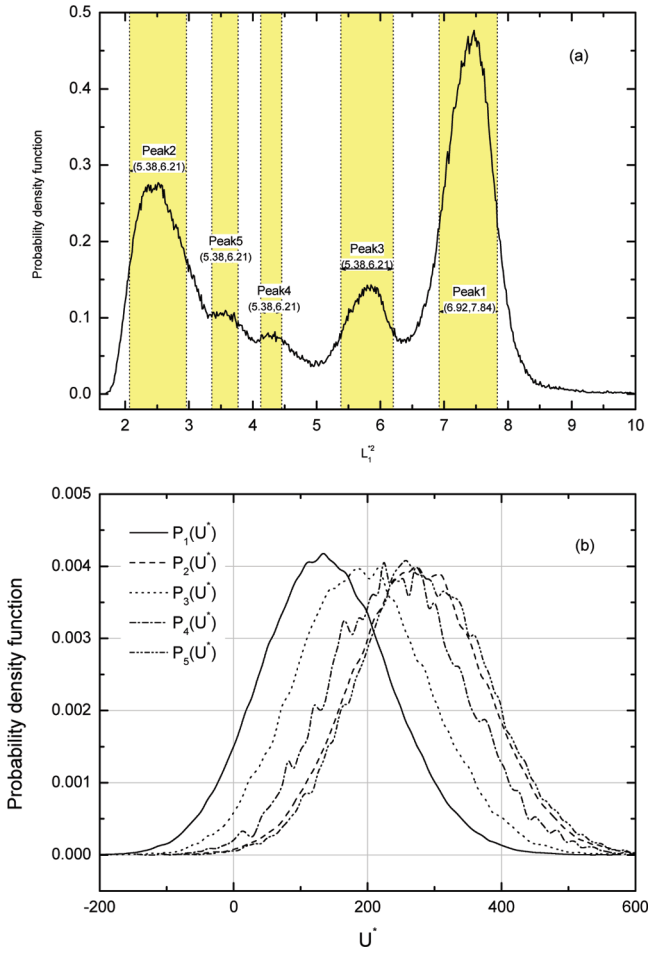



***Figure 9.*** (a) Probability density function of $L_1^{*2}$ and the definition of the configuration groups based on the peak locations. (b) The normalized energy distributions for the chain configurations in each configuration group.

and 2, with the peak with smaller $L_1^{*2}$ having a higher energy distribution $P_i(U^*)$. However, peak 5 is an exception to the relationship between $L_1^{*2}$ and $U^*$. It is located at higher $L_1^{*2}$ than peak 2 but also produces an energy distribution $P_5(U^*)$ at slightly higher energy than $P_2(U^*)$.

Before we show the relation between chain dimension and shape, we would like to comment on the use of a Landau free energy expression $A_L(Q)$ in dependence on a structure parameter $Q$, viz, $A_L(Q) = A - kT \log(p_Q(Q))$, where $A$ is the Helmholtz free energy and $p_Q(Q)$ the canonical probability distribution for an order parameter $Q$. Bimodal and multimodal probability distributions of $R_g^{*2}$ or $L_1^{*2}$ as have been presented before in single chain studies were used to obtain a Landau free energy expression.[91,96] Multi(bi)modality in $\langle L_1^{*2} \rangle$ leads to a multi(bi)modal Landau free energy implying true thermodynamic coexistence between different conformational states. However, in our case, this conclusion is not valid, as the potential energy probability density distribution is unimodal (see Figure 8a) and not multi- or bimodal. Only when there is a one-to-one correspondence between the energies of the different conformational states and the structure parameter $Q$ (e.g., $R_g^{*2}$ or $L_1^{*2}$) characterizing the different conformational states, bi(multi)modality in the energy probability density (necessary for thermodynamic coexistence) would lead to a bi(multi)modality in the conformational parameter that could be correctly interpreted in terms of thermodynamic coexistence. Hence, the bi- or multimodality in $\langle L_1^{*2} \rangle$ or $\langle R_g^{*2} \rangle$ observed here does not mean true coexistence as in a first order phase transition. As discussed before, for the case studied here, the solid−liquid transition is a continuous transition.[86] Therefore, the use and applicability of a Landau free energy expression requires that a unique one-to-one correspondence exists between the order parameter probability density used in the Landau free energy approach and the energy probability density relevant for thermodynamic coexistence behavior. This condition should be specifically checked to ascertain the applicability of the Landau free energy approach.

**6. Instantaneous Conformational Shapes and Shape Parameters at $T_m^{0*}$.** The pattern found in Figure 8b inspired us to analyze the conformational data using shape parameters also used by Binder et al.[87,101] Two instantaneous shape parameters $K_1$ and $K_2$ are defined by

$$K_1 = \frac{L_2^{*2} + L_3^{*2}}{L_1^{*2} + L_2^{*2}}, K_2 = \frac{L_1^{*2} + L_3^{*2}}{L_1^{*2} + L_2^{*2}} \quad (3)$$

with $L_1^{*2} \geq L_2^{*2} \geq L_3^{*2}$ being the instantaneous principal components of the radius of gyration. Note that $K_2 \geq K_1$ *and* $K_2 \geq 0.5$ are always true. As their name suggests, these shape parameters can describe many different shapes of chain molecules. For example, the polymer chain has a spherical shape when $K_1 \rightarrow 1$ and $K_2 \rightarrow 1$, a thin rod conformation when $K_1 \rightarrow 0$ and $K_2 \rightarrow 1$, or a thin round disk or ring-like structure when $K_1 \rightarrow 0.5$ and $K_2 \rightarrow 0.5$. Other combinations of $K_1$ and $K_2$ represent intermediate chain shapes in between those extremes, e.g., ellipsoid.

$K_1$ and $K_2$ for the $5 \times 10^5$ instantaneous chain configurations are calculated, and the contour map of the resulting population as a function of the shape parameters $K_1$ and $K_2$ is presented in Figure 10. According to the definitions of the shape parameters, $K_2 \geq 0.5$ and $K_2 > K_1$. As we can see at $T_m^{0*}$, the population covers almost all of the possible combinations of $K_1$ and $K_2$. This is a very interesting result, since the broad distribution in the shape parameter space implies that the chain adopts not only the globular and rod-like conformations but also very diverse shapes at the transition temperature. Furthermore, the contour map shows several highly populated regions, which suggests that some shapes are particularly more favorable. The most dense population is located at very small $K_1$ and large $K_2$ values, which define the rod-like structure in agreement with the folded chain lamellar shape. Another highly populated region in Figure 10a is centered at $K_1 \approx 0.65$ and $K_2 \approx 0.85$, which are the values typical for ellipsoidal shapes. It is clear that this region is broader than the region of the rod-like shapes, which implies that the shapes of the individual configurations vary a lot (from more disk-like to more spherical ellipsoids). Besides these two most obvious regions, there are also some less populated but recognizable high density regions in between, as can be seen in Figure 10a and b.

With the same definition for the peaks as given in Figure 9a, chain configurations having similar $L_1^{*2}$ are grouped together, and their shape parameters are retrieved to locate

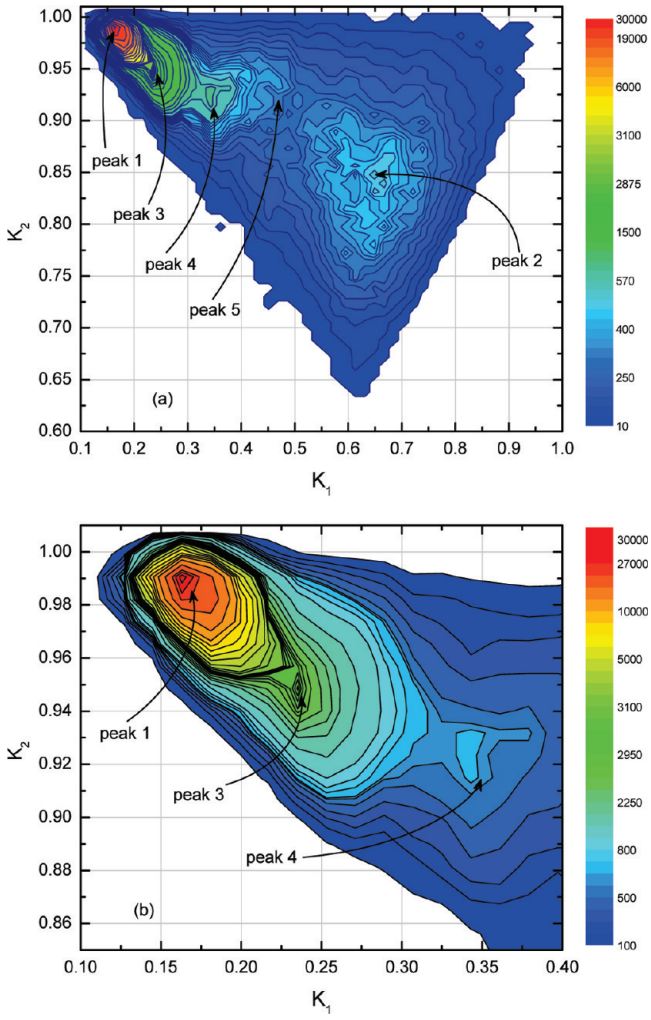



***Figure 10.*** (a) Contour map of the population of the $5 \times 10^5$ chain configurations in the 2D space defined by the shape parameters $K_1$ and $K_2$ at the equilibrium melting temperature $T_m^{0*} \cong 4.234$ and (b) detail of the contour map in the small $K_1$ region.

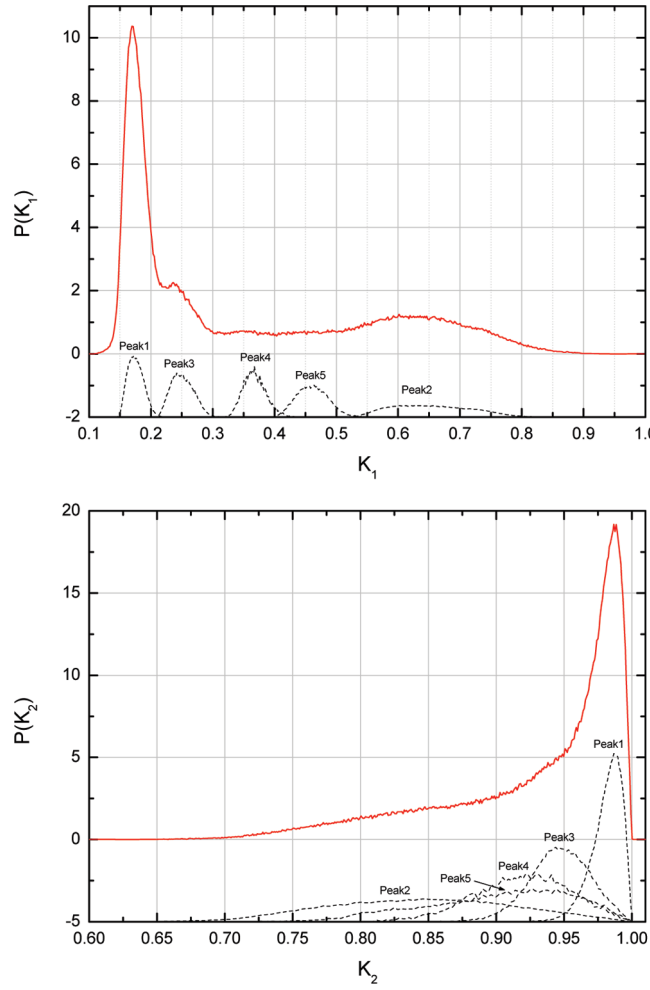



***Figure 11.*** Probability density distribution of the $5 \times 10^5$ chain configurations as a function of shape parameter $K_1$ (a) and $K_2$ (b) at the equilibrium melting temperature $T_m^{0*} \cong 4.234$. The normalized distributions of $K_1$ and $K_2$ of the chain configurations in the separate configuration groups are also presented in a and b, respectively.

them on the shape parameter map shown in Figure 10. In Figure 10a, the five configuration groups are indicated by the arrows in the $K_1-K_2$ shape parameter map. Then, the shape parameter distributions of all the configurations and the configurations belonging to the defined groups are presented in Figure 11a and b for $K_1$ and $K_2$, respectively. Along the $K_1$ axis, the five distributions are totally separated, whereas along $K_2$ they strongly overlap. This is due to the fact that the ratio between the largest component $L_1^{*2}$ and two smaller components $L_2^{*2}$ or $L_3^{*2}$ of the radius of gyration is more prominent than the ratio between the two smaller ones $L_2^{*2}/L_3^{*2}$. As a result, $K_1$ is a better measure for identifying the different individual chain shapes at the equilibrium melting temperature. $P(K_1)$ shows obvious multiple peak features as also observed in $P(L_1^{*2})$.

**7. Instantaneous Shapes at $T_m^{0*}$: Snapshots of Representative Chain Conformations.** The distributions of the shape parameters $K_1$ and $K_2$ give statistical information. From Figure 11, we know peak 1 includes chain configurations having the smallest $K_1 \rightarrow 0$ but largest $K_2 \rightarrow 1$, which clearly tells us most of them are rod-shaped. Peak 2 (both small $K_1$ and $K_2$) collects conformations which are more spherical. Peaks 3, 4, and 5 sit in between, which means the overall shapes change from a perfect rod to ellipsoid.

Molecular simulations allow us to study individual configurations in great detail. To verify our analysis based on the shape parameters and to characterize the chain configurations in the aforementioned configuration groups, snapshots for the chain configurations in each group are extracted and visualized.

First, the extreme configuration at the smallest $K_1 = 0.114$ and the largest $K_2 = 0.992$ observed during the simulation is shown in Figure 12. The polymer chain folds itself into six parallel stems forming a long and thin rod. However, the parallel stems are not perfectly aligned, and this causes the $CH_2$ units in the loops and at the ends of the parallel stems to have fewer van der Waals contacts with neighboring stems than when the stems would be properly aligned. As a consequence, the potential energy increases. This explains why this configuration does not fall into the most probable region defined by peak 1 in the shape parameter map.

The second extreme is located in the upper-right corner of the shape parameter map triangle with both $K_1$ and $K_2$

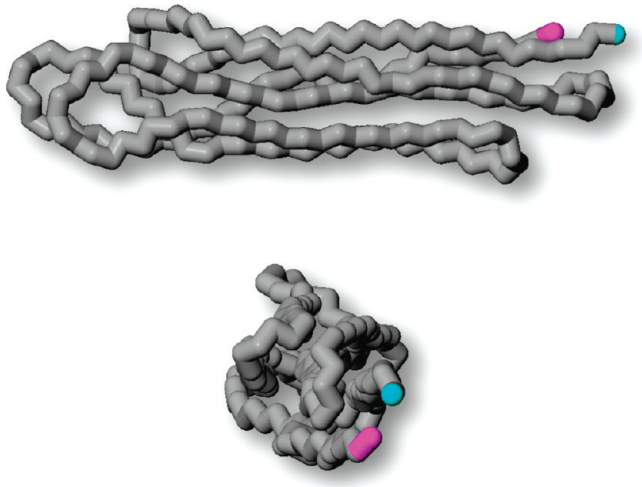


***Figure 12.*** Lateral view and intersection view of the chain configuration ($N = 200$) with shape parameters $K_1 = 0.114$ and $K_2 = 0.992$ at $T_m^{0*} \cong 4.234$.

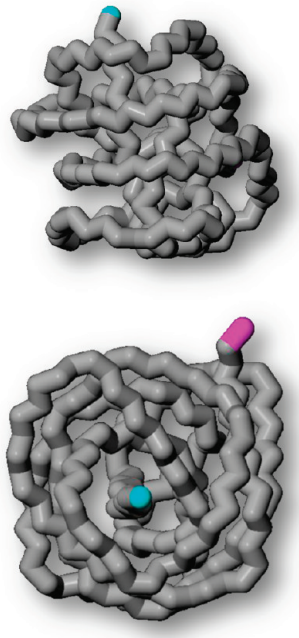


***Figure 13.*** Lateral view and intersection view of the chain configuration ($N = 200$) with shape parameters $K_1 = 0.876$ and $K_2 = 0.930$ at $T_m^{0*} \cong 4.234$.

approaching 1 ($K_1 = K_2 = 1$ represent a perfect sphere). In our simulation, the configuration with the largest $K_1$ and the largest $K_2$ simultaneously is found at $K_1 = 0.876$ and $K_2 = 0.930$ and is shown in Figure 13. The side view and the top view show that the chain has similar sizes in all dimensions, although it is far from a perfect sphere. Furthermore, the chain shows some kind of ordering: some short *all-trans* segments swirl about one axis and keep parallel to each other.

Disk-like shapes have similar $K_1$ and $K_2$ values. A chain configuration having $K_1$ and $K_2$ the closest to each other is located at $K_1 = 0.601$ and $K_2 = 0.638$ and is shown in Figure 14. It looks more like a "donut" than a disk. Similar toroidal structures for a polymer chain have also been observed in other simulations[24,87,96,97,99,101,102] for polymer chains with

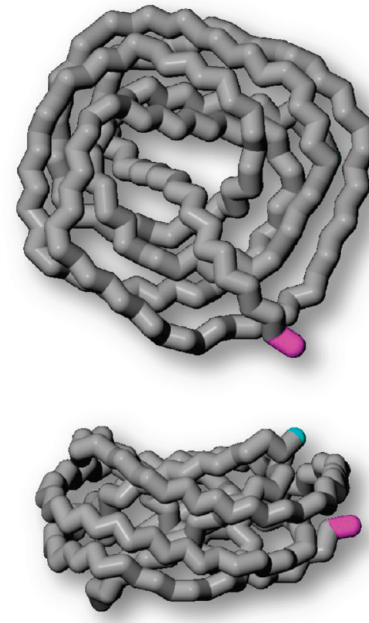


***Figure 14.*** Lateral view and intersection view of the chain configuration ($N = 200$) with shape parameters $K_1 = 0.601$ and $K_2 = 0.638$ at $T_m^{0*} \cong 4.234$.

a high degree of stiffness. Although polyethylene is normally categorized as a flexible chain, toroidal shapes can also be explored by the PE chain.

The three extreme conformations presented above do not represent the most probable conformation in each group. However, they help us to understand or imagine what an actual chain configuration looks like and the link between the shape parameters and the chain unit arrangement. To study the most probable configurations, we show in Figure 15 the snapshots of representative chain configurations located in the peak positions of the five different groups in Figure 11a and b.

Figure 15a shows a typical rod-like structure formed by parallel arranged *all-trans* stems found at $K_1 = 0.172$ and $K_2 = 0.987$. Unlike the configuration shown in Figure 12, the *all-trans* stems appear to be better aligned and form a well ordered rod with both ends capped by the chain loops. There are seven folds in the rod-shaped configuration, which is less than the number found for the conformation in Figure 12, and therefore the rod is shorter. This configuration has lower energy since both the conformational energy and the nonbonded interaction are minimized.

For the other highly populated peak, peak 2, the favorable configuration at $K_1 = 0.602$ and $K_2 = 0.864$ looks like the one presented in Figure 15b, a compact ellipsoid with a certain degree of ordering. The configuration presented in Figure 15c is found in peak 3 (the one next to group 1 along $K_1$) with $K_1 = 0.242$ and $K_2 = 0.944$. It has eight folds and even shorter *all-trans* segments. It is also twisted because of the occurrence of a few defects (i.e., a few *gauche* conformations) in the stems. Figure 15d and e are the configurations at the centers of peaks 4 and 5 in the shape parameter map, respectively. These groups of intermediate $K_1$ and $K_2$ collect many different shapes which are located

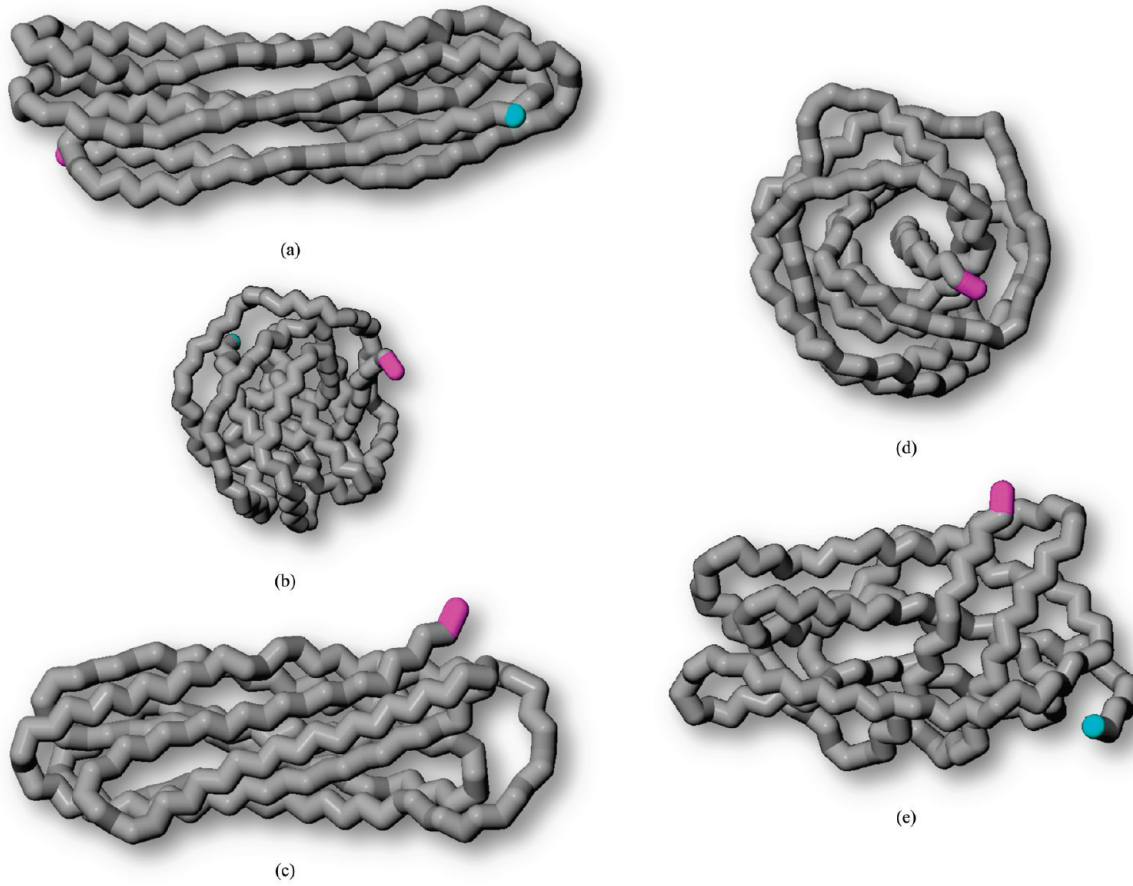



***Figure 15.*** Snapshots of typical chain configurations ($N = 200$) selected in the peak maxima of the shape parameter distributions shown in Figure 11. (a) $K_1 = 0.172$ and $K_2 = 0.987$, (b) $K_1 = 0.602$ and $K_2 = 0.864$, (c) $K_1 = 0.242$ and $K_2 = 0.944$, (d) $K_1 = 0.367$ and $K_2 = 0.930$, and (e) $K_1 = 0.463$ and $K_2 = 0.926$.

between the perfect rod and the ellipsoid. Figure 15d presents the configuration found at $K_1 = 0.367$ and $K_2 = 0.930$: it is a thicker toroid. The configuration in Figure 15e with $K_1 = 0.463$ and $K_2 = 0.926$ looks like an intermediate state between the folded rod and the toroidal configuration with the *all-trans* stems tilted.

## Conclusions

Using the parallel tempering canonical molecular dynamics method, a polyethylene chain of 200 $CH_2$ united atom units is simulated in a very broad temperature range covering both the coil−globule and the globule−rod transitions of the polymer chain. The structural and conformational properties of the polyethylene chain derived from the simulation results are discussed.

The radial distribution function and the structure factor of the chain have been analyzed in great detail. The peaks in the RDF have been assigned to characteristic distances in the chain: bond length, bond angle, torsional distances of four $CH_2$ units up to 10 $CH_2$ units provide distinct and sharp peaks in the RDF which gradually broaden with increasing temperature. The peaks related to torsional distances of four and more $CH_2$ units gradually disappear with increasing temperature. However, these peaks do not disappear abruptly at the equilibrium melting temperature, and therefore these peaks are not typical for the crystalline state or the equilibrium transition temperature. At the equilibrium melting temperature, a broad peak appears that becomes more pronounced at lower temperatures. This broader peak stems from next nearest neighbor hexagonal distances characteristic of the ordered crystalline lamellar state.

Also in the structure factor $S(q^*)$, the different peaks have been analyzed and assigned to different length scales in the chain molecule. At very high $q^*$, $S(q^*)$ is related to smaller length scales, and distinct features due to bond lengths, bond angles, and torsion angles were resolved and can serve as fingerprints for these atomistic distances in the chain. At smaller $q^*$ values, in the regime $[\sqrt{(\langle R_g^{*2}\rangle)}]^{-1} \ll q^* \ll l_0^{*-1}$, the structure factor reflects both local and larger-scale structure information. In this regime, two peaks emerge as the temperature decreases that are located at the lowest temperature $T^*_{min} = 2.52$ at $q^* \cong 3.189$ and $q^* \cong 5.186$. The peak at $q^* \cong 5.186$ is also present in the globular state and is thus not representative for the crystallized chain but is related to nearest neighbor distances which are typical for the high segment densities in the crystalline and globular states. On the other hand, the peak at $q^* \cong 3.189$ is related to the formation of periodic structures and has successfully been used to determine the equilibrium melting temperature on the basis of structure factor (scattering) information. The use of $S(q^*)$ in this manner links to the experimental practice of studying crystallization/melting transitions in materials

using scattering techniques, and the equilibrium melting temperature established in this way is in quantitative agreement with the equilibrium melting temperature determined from other properties.[86]

Following Hammouda, the full $q^*$-range scaling behavior of $S(q^*)$ is studied in detail and related to the conformational properties. At low temperatures, the scaling analysis indicates that the chain is in a rod-like structure. At temperatures slightly higher than the equilibrium melting temperature, the scaling behavior is found to be in agreement with the globular state, and at the highest temperature, the scaling behavior is indicative of the expanded excluded volume coil.

In the Θ region, the ideal chain behavior is not only shown by the scaling analysis but is also obtained from a quantitative comparison of the full $S(q^*)$ with the structure factor of the equivalent Kuhn chain. In doing this detailed analysis, we have been able to determine the Θ region of our chain with good accuracy without having to resort to the studies of different chain lengths.

However, we also have shown that a correct assignment of the Θ region requires care, and the presence of a horizontal line piece in the Kratky plot is not sufficient evidence of Θ conditions. For our PE 200 chain, the equivalent Kuhn chain only contains ca. 30 Kuhn segments, which is not long enough to fulfill the long or Gaussian chain approximations. As a result, in the Kratky plot, the ideal chain behavior is not reflected by a horizontal line piece at intermediate $q^*$ in $S(q^*)\cdot q^{*2}$ vs $q^*$ (typical for Gaussian chains and in practice used as the signature of ideal chain behavior) but has a sigmoidal shape. However, at higher temperatures when the chain behaves already as an excluded volume chain, a horizontal line piece is present in the Kratky plot, and one could mistakenly take this temperature as the Θ temperature. Therefore, for smaller chain lengths, only a comparison of the full structure factor leads to consistent and accurate determination of the Θ region.

A detailed analysis of the chain conformational shapes was done at the equilibrium melting temperature. The conformational shapes of the PE chain were represented in conformational distribution functions using the eigenvalue of the largest principal component of the radius of gyration tensor as well as the shape parameters $K_1$ and $K_2$ as order parameters. The conformational distribution functions have a multimodal dependence on the selected order parameters. The different distinct conformational regions that were found include the already know expanded coil, globular, and chain folded lamellar structures. However, also a distinct region of toroidal or more disk-like structures is discernible. These structures are normally linked to stiffer chain molecules and are rather unexpected for the flexible PE molecule. So far, the toroidal state had not been identified in simulations on PE chains. From our results, it is now clear that, at the equilibrium transition temperature, the PE chain can take, next to the common rod-like lamellar and globular structures, indicative of the solid and liquid states, also a greater diversity of conformational shapes, including toroidal shapes.

Although different conformational regions are found, their potential energy probability densities are not distinct but largely overlap. As a consequence, the use of the multimodal conformational density distributions in a Landau free energy lead to erroneous predictions of true thermodynamic coexistence behavior between different conformational states. In our case studied here, the chain conformations can continuously change from one region to the other, resulting also in a continuous transition instead of a first-order transition characterized by coexisting states. The applicability of the Landau free energy approach should always be checked and is only valid when there is a one to one correspondence between the energy and conformational distributions. The shape parameter analysis is further clarified and exemplified by a study of characteristic snapshots from the conformational distribution function.

**Acknowledgment.** The authors thank the Funds for Scientific Research Flanders for financial support. T.L. is indebted to the Katholieke Universiteit Leuven for a postdoctoral fellowship (OT 03/93). The computations for this research have been done on the VIC HPC supercomputer of the K. U. Leuven.

**Supporting Information Available:** Model description and force field parameters and information on the performance and statistics of the REMD simulation. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hammouda, B. *J. Appl. Crystallogr.* **2010**, *43*, 716–719.

(2) Xu, J.; Zhu, Z.; Luo, S.; Wu, C.; Liu, S. *Phys. Rev. Lett.* **2006**, *96*.

(3) Wang, X.; Wu, C. *Macromolecules* **1999**, *32*, 4299–4301.

(4) Wang, X.; Qiu, X.; Wu, C. *Macromolecules* **1998**, *31*, 2972–2976.

(5) Wu, C.; Zhou, S. *Phys. Rev. Lett.* **1996**, *77*, 3053.

(6) Chu, B.; Ying, Q.; Grosberg, A. Y. *Macromolecules* **1995**, *28*, 180–189.

(7) Swislow, G.; Sun, S.; Nishio, I.; Tanaka, T. *Phys. Rev. Lett.* **1980**, *44*, 796.

(8) Huser, T.; Yan, M.; Rothberg, L. J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11187–11191.

(9) Tanaka, T. *Polymer* **1979**, *20*, 1404–1412.

(10) Perkins, T.; Smith, D.; Chu, S. *Science* **1994**, *264*, 819–822.

(11) Deniz, A. A.; Mukhopadhyay, S.; Lemke, E. A. *J. R. Soc. Interface* **2008**, *5*, 15–45.

(12) Perkins, T.; Smith, D.; Larson, R.; Chu, S. *Science* **1995**, *268*, 83–87.

(13) Kumaki, J.; Nishikawa, Y.; Hashimoto, T. *J. Am. Chem. Soc.* **1996**, *118*, 3321–3322.

(14) Kumaki, J.; Hashimoto, T. *J. Am. Chem. Soc.* **2003**, *125*, 4907–4917.

(15) Ortiz, C.; Hadziioannou, G. *Macromolecules* **1999**, *32*, 780–787.

(16) Bemis, J. E.; Akhremitchev, B. B.; Walker, G. C. *Langmuir* **1999**, *15*, 2799–2805.

(17) Rief, M.; Oesterhelt, F.; Heymann, B.; Gaub, H. E. *Science* **1997**, *275*, 1295–1297.

(18) Zlatanova, J.; Lindsay, S. M.; Leuba, S. H. *Prog. Biophys. Mol. Biol.* , *74*, 37–61.

(19) Fritz, J.; Anselmetti, D.; Jarchow, J.; Fernandez-Busquets, X. *J. Struct. Biol.* **1997**, *119*, 165–171.

(20) Stockmayer, W. H. *Die Makromolekulare Chemie* **1960**, *35*, 54–74.

(21) Mattice, W. L.; Suter, U. W. *Conformational Theory of Large Molecules: The Rotational Isomeric State Model in Macromolecular Systems*; 1st ed.; Wiley-Interscience, 1994.

(22) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Reprint.; Oxford Univ Pr (Sd), 1989.

(23) Yamakawa, H. *Modern theory of polymer solutions*; Harper & Row, 1971.

(24) Binder, K.; Paul, W.; Strauch, T.; Rampf, F.; Ivanov, V.; Luettmer-Strathmann, J. *J. Phys.: Condens. Matter* **2008**, *20*, 494215.

(25) Liao, Q.; Jin, X. *J. Chem. Phys.* **1999**, *110*, 8835.

(26) Baumgärtner, A. *J. Chem. Phys.* **1980**, *72*, 871.

(27) Jeppesen, C.; Kremer, K. *Europhys. Lett.* **1996**, *34*, 563–568.

(28) Hu, W. *J. Chem. Phys.* **1998**, *109*, 3686.

(29) Wittkop, M.; Kreitmeier, S.; Göritz, D. *J. Chem. Phys.* **1996**, *104*, 3373.

(30) Noguchi, H.; Yoshikawa, K. *J. Chem. Phys.* **1998**, *109*, 5070.

(31) Takahashi, M.; Yoshikawa, K.; Vasilevskaya, V. V.; Khokhlov, A. R. *J. Phys. Chem. B* **1997**, *101*, 9396–9401.

(32) Taylor, M. P.; Paul, W.; Binder, K. *J. Chem. Phys.* **2009**, *131*, 114907.

(33) Narambuena, C.; Leiva, E.; Chávez-Páez, M.; Pérez, E. *Polymer* **2010**, *51*, 3293–3302.

(34) Ivanov, V. A.; Stukan, M. R.; Müller, M.; Paul, W.; Binder, K. *J. Chem. Phys.* **2003**, *118*, 10333.

(35) Stukan, M. R.; Ivanov, V. A.; Grosberg, A. Y.; Paul, W.; Binder, K. *J. Chem. Phys.* **2003**, *118*, 3392.

(36) Kremer, K.; Binder, K. *Comput. Phys. Rep.* **1988**, *7*, 259–310.

(37) Zhou, Y.; Hall, C. K.; Karplus, M. *Phys. Rev. Lett.* **1996**, *77*, 2822–2825.

(38) Liang, H.; Chen, H. *J. Chem. Phys.* **2000**, *113*, 4469.

(39) Baysal, B. M.; Karasz, F. E. *Macromol. Theory Simul.* **2003**, *12*, 627–646.

(40) Swislow, G.; Sun, S.; Nishio, I.; Tanaka, T. *Phys. Rev. Lett.* **1980**, *44*, 796–798.

(41) Harano, Y.; Kinoshita, M. *J. Phys.: Condens. Matter* **2006**, *18*, L107–L113.

(42) Polson, J. M.; Moore, N. E. *J. Chem. Phys.* **2005**, *122*, 024905.

(43) Kunugi, S.; Tada, T.; Yamazaki, Y.; Yamamoto, K.; Akashi, M. *Langmuir* **2000**, *16*, 2042–2044.

(44) Martemyanova, J. A.; Stukan, M. R.; Ivanov, V. A.; Müller, M.; Paul, W.; Binder, K. *J. Chem. Phys.* **2005**, *122*, 174907.

(45) Liao, Q.; Dobrynin, A. V.; Rubinstein, M. *Macromolecules* **2006**, *39*, 1920–1938.

(46) Noguchi, H.; Yoshikawa, K. *J. Chem. Phys.* **1998**, *109*, 5070.

(47) Berghmans, H.; Deberdt, F. *Philos. Trans. R. Soc. Lond., Ser. A* **1994**, *348*, 117–128.

(48) Rampf, F.; Binder, K.; Paul, W. *J. Polym. Sci., Part B: Polym. Phys.* **2006**, *44*, 2542–2555.

(49) Baysal, B. M.; Karasz, F. E. *Macromol. Theory Simul.* **2003**, *12*, 627–646.

(50) Moore, M. A. *J. Phys. A: Math. Gen.* **1977**, *10*, 305.

(51) Stockmayer, W. H. *Makromol. Chem.* **1960**, *35*, 54–74.

(52) Muthukumar, M. In *Interphases and Mesophases in Polymer Crystallization III*; 2005; pp 241−274.

(53) Binder, K.; Baschnagel, J.; Müller, M.; Paul, W.; Rampf, F. *Macromol. Symp.* **2006**, *237*, 128–138.

(54) Zhou, Hall; Karplus, *Phys. Rev. Lett.* **1996**, *77*, 2822–2825.

(55) Larini, L.; Barbieri, A.; Prevosto, D.; Rolla, P. A.; Leporini, D. *J. Phys.: Condens. Matter* **2005**, *17*, L199.

(56) Muthukumar, M. *In Progress in Understanding of Polymer Crystallization;* **2007**, 1–18.

(57) Doye, J. P. K.; Sear, R. P.; Frenkel, D. *J. Chem. Phys.* **1998**, *108*, 2134–2142.

(58) Chen, C. M.; Higgs, P. G. *J. Chem. Phys.* **1998**, *108*, 4305–4314.

(59) Swendsen, R. H.; Wang, J. *Phys. Rev. Lett.* **1986**, *57*, 2607.

(60) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(61) Geyer, C. J.; Keramidas, E. M. Keramidas, EM, Ed 156−163.

(62) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042.

(63) Wang, J. S.; Swendsen, R. H. *Progress of Theoretical Physics-Supplement* **2005**, 317–323.

(64) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(65) Hukushima, K.; Takayama, H.; Nemoto, K. *Int. J. Mod Phys C* **1996**, *7*, 337–344.

(66) Hansmann, U. H. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(67) Tesi, M. C.; Rensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82*, 155–181.

(68) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249–253.

(69) Berg, Neuhaus *Phys. Rev. Lett.* **1992**, *68*, 9–12.

(70) Berg, B. A.; Celik, T. *Phys. Rev. Lett.* **1992**, *69*, 2292.

(71) Paul, W.; Müller, M. *J. Chem. Phys.* **2001**, *115*, 630.

(72) Paul, W.; Müller, M. *Comput. Phys. Commun.* **2002**, *146*, 113–117.

(73) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.

(74) Ryckaert, J. P.; Bellemans, A. *Faraday Discuss. Chem. Soc.* **1978**, *66*, 95–106.

(75) Sundararajan, P. R.; Kavassalis, T. A. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 2541–2549.

(76) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(77) Tobias, D. J.; Martyna, G. J.; Klein, M. L. *J. Phys. Chem.* **1993**, *97*, 12959–12966.

(78) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.

(79) Katzgraber, H. G.; Trebst, S.; Huse, D. A.; Troyer, M. *J. Stat. Mech.* **2006**, *2006*, P03018–P03018.

(80) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *76*, 057102.

(81) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *76*, 065701.

(82) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386–10387.

(83) Sindhikara, D. J.; Emerson, D. J.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 2804–2808.

(84) Kone, A.; Kofke, D. A. *J. Chem. Phys.* **2005**, *122*, 206101.

(85) Rosta, E.; Hummer, G. *J. Chem. Phys.* **2009**, *131*, 165102.

(86) Li, T.; Jiang, Z.; Yan, D.; Nies, E. *Polymer* **2010**, *51*, 5612–5622.

(87) Ivanov, V. A.; Paul, W.; Binder, K. *J. Chem. Phys.* **1998**, *109*, 5659.

(88) Higgins, J. *Polymers and neutron scattering*; Clarendon Press: Oxford, 1996.

(89) Destree, M.; Lyulin, A.; Ryckaert, J. *Macromolecules* **1996**, *29*, 1721–1727.

(90) Sundararajan, P. R.; Kavassalis, T. A. *Faraday Trans.* **1995**, *91*, 2541.

(91) Doye, J. P. K.; Sear, R. P.; Frenkel, D. *J. Chem. Phys.* **1998**, *108*, 2134.

(92) Noguchi, H.; Yoshikawa, K. *J. Chem. Phys.* **1998**, *109*, 5070.

(93) Yoshikawa, K.; Yoshinaga, N. *J. Phys.: Condens. Matter* **2005**, *17*, S2817–S2823.

(94) Odijk, T. *Macromolecules* **1993**, *26*, 6897–6902.

(95) Cifra, P.; Benková, Z.; Bleha, T. *J. Phys. Chem. B* **2008**, *112*, 1367–1375.

(96) Sarraguça, J. M. G.; Dias, R. S.; Pais, A. A. C. C. *J Biol Phys* **2006**, *32*, 421–434.

(97) Stukan, M. R.; Ivanov, V. A.; Grosberg, A. Y.; Paul, W.; Binder, K. *J. Chem. Phys.* **2003**, *118*, 3392.

(98) Hud, N. V.; Vilfan, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 295–318.

(99) Stukan, M.; An, E.; Ivanov, V.; Vinogradova, O. *Phys. Rev. E* **2006**, *73*.

(100) Vasilevskaya, V. V.; Khokhlov, A. R.; Kidoaki, S.; Yoshikawa, K. *Biopolymers* **1997**, *41*, 51–60.

(101) Ivanov, V. A.; Stukan, M. R.; Vasilevskaya, V. V.; Paul, W.; Binder, K. *Macromol. Theory Simul.* **2000**, *9*, 488–499.

(102) Narambuena, C.; Leiva, E.; Chávez-Páez, M.; Pérez, E. *Polymer* **2010**, *51*, 3293–3302.

CT100513Y

JCTC Journal of Chemical Theory and Computation

# TMSmesh: A Robust Method for Molecular Surface Mesh Generation Using a Trace Technique

Minxin Chen*,† and Benzhuo Lu*,‡

*Department of Mathematics, SooChow University, Suzhou 215006, China and State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Received July 4, 2010

**Abstract:** Qualified, stable, and efficient molecular surface meshing appears to be necessitated by recent developments for realistic mathematical modeling and numerical simulation of biomolecules, especially in implicit solvent modeling (e.g., see a review in B. Z. Lu et al. *Commun. Comput. Phys.* **2008**, *3*, 973−1009). In this paper, we present a new method: tracing molecular surface for meshing (TMSmesh) the Gaussian surface of biomolecules. The method computes the surface points by solving a nonlinear equation directly, polygonizes by connecting surface points through a trace technique, and finally outputs a triangulated mesh. TMSmesh has a linear complexity with respect to the number of atoms and is shown to be capable of handling molecules consisting of more than one million atoms, which is usually difficult for the existing methods for surface generation used in molecular visualization and geometry analysis. Moreover, the meshes generated by TMSmesh are successfully tested in boundary element solutions of the Poisson−Boltzmann equation, which directly gives rise to a route to simulate electrostatic solvation of large-scale molecular systems. The binary version of TMSmesh and a set of representative PQR benchmark molecules are downloadable at our Web page http://lsec.cc.ac.cn/∼lubz/Meshing.html.

## Introduction

Molecular surface mesh is widely used for visualization and geometry analysis in computational structural biology and structural bioinformatics. Recent developments in realistic mathematical modeling and numerical simulation of biomolecular systems raise new demands for qualified, stable, and efficient surface meshing, especially in implicit-solvent modeling (e.g., see a review in ref 1). Main concerns for improvement on existing methods for molecular surface mesh generation are efficiency, robustness, and mesh quality. Efficiency is necessary for simulations/computations requiring frequent mesh generation or requiring mesh of large systems. Robustness here means the meshing method is stable and can treat various, even arbitrary, sizes of molecular systems within computer power limitations. Mesh quality relates to mesh smoothness (avoiding sharp solid angles, etc.), uniformness (avoiding elements with very sharp angles or zero area), and topological correctness (avoiding isolated vertices, element intersection, single-element-connected edges, etc.). The quality requirement is critical for some numerical techniques, such as finite element method, to achieve converged and reasonable results, which makes it a more demanding task in this respect than the mesh generations only for the purposes of visualization or some structural geometry analysis.

Various definitions of molecular surface, including the van der Waals (VDW) surface, solvent accessible surface (SAS), solvent excluded surface (SES), molecular skin surface,[2] minimal molecular surface,[3] and Gaussian surface, etc., have been proposed to describe the shapes of molecular structure. The VDW surface is defined as the surface of the union of the spherical atomic surfaces with the VDW radius of each atom in the molecule. The SAS and SES are represented by

* Corresponding authors. E-mail: chenmx@gmail.com, bzlu@lsec.cc.ac.cn.

† SooChow University.

‡ Chinese Academy of Sciences.

10.1021/ct100376g 
Published on Web 11/30/2010

the trajectory of the center and the interboundary of a rolling probe on the VDW surface, respectively. The molecular skin surface is the envelope of an infinite family of spheres derived from atoms by convex combination and shrinking. The minimal molecular surface is defined as a result of the surface free energy minimization. Different from these definitions, the Gaussian surface is defined as a level set of the summation of the Gaussian kernel functions as follows

$$\{\vec{x} \in \mathbb{R}^3, \phi(\vec{x}) = t_0\} \tag{1}$$

where

$$\phi(\vec{x}) = \sum_{i=1}^{N} e^{d(||\vec{x} - \vec{c}_i||^2/r_i^2 - 1)} \tag{2}$$

$\vec{c}_i$ and $r_i$ are the location and radius of atom $i$, the parameter $d$ is negative and controls the decay speed of the kernel functions. When $|d|$ is increased, the resulting Gaussian surface is closer to the VDW surface. In this work, the value of $d$ and $t_0$ are set as $-1$ and 1, respectively. Compared with other definitions of molecular surface, the Gaussian surface is smooth and more suitable to represent the electron density of a molecule.[4] The VDW surface, SAS, and SES can be approximated well by the Gaussian surface with proper parameter selection.[4,5] The Gaussian surface has been widely used in many problems in computational biology, such as docking problems,[6] molecular shape comparisons,[7] calculating SAS areas,[8] and the generalized Born models.[9]

With various definitions of molecular surface that have been proposed, numerous works have been devoted to the computation of molecular surface. The representative ones are described as follows. In 1983, Connolly proposed algorithms to calculate the molecular surface and SAS analytically.[10,11] In 1995, a popular program, GRASP, for visualizing molecular surfaces was presented.[12] In 1997, Vorobjev et al. proposed SIMS, a method of calculating a smooth invariant molecular dot surface, in which an exact method for removing self-intersecting parts and smoothing the singular regions of the SES was presented.[13] Sanner et al. presented a tool based on α shapes,[14] named MSMS, for meshing the SES.[15] Ryu et al. proposed a method based on $\beta$ shapes that are a generalization of α shapes.[16] More recently, Zhang et al. used a modified dual contouring method to generate mesh for biomolecular structures,[17] and a later tool, GAMer, was developed for improving the mesh quality.[18] Can et al. proposed LSMS to generate the SES on grid points using level-set methods.[19] Chavent et al. presented MetaMol to visualize the molecular skin surface using the ray-casting method,[20] and Cheng et al. used restricted union of balls to generate mesh for molecular skin surfaces.[21] So far, these methods or tools usually successfully calculated different surfaces of small- or medium-sized biomolecules, but they are not suitable for large molecules with more than hundreds of thousands of atoms. Moreover, most of these methods, such as GRASP, MSMS, and LSMS were designed for molecular visualization and geometry analysis in computational structure biology or structural bioinformatics. Among those, MSMS is the most widely used one for molecular surface triangulation because of its high efficiency. However, the generated mesh is not a manifold and is composed of very irregular triangles. For some numerical modeling using, for instance, finite element/boundary element methods, the mesh quality usually needs to be improved through mesh topology checking (picking out the irregular nodes/edges/elements and rearranging the mesh), surface mesh smoothing, and so on.[1] In this paper, we develop a robust method, named TMSmesh that is capable of finishing meshing the Gaussian surface for biomolecules consisting of more than one million atoms in 30 min on a typical 2010 PC, and the mesh quality is shown to be applicable to boundary element method simulations of biomolecular electrostatics.

As the Gaussian surface is an implicit surface, the existing techniques for triangulating implicit surface can be used for the Gaussian surface. These methods are divided into two main categories: spatial partition and continuation methods. The well-known marching cubes[22] and dual contouring methods[23] are examples of the spatial partition methods. This kind of method divides the space into cells and polygonizes the implicit surface in the cell whose vertices have different signs of the implicit function. An assumption is required that the implicit function is linear in the cell. As shown in the following sections, TMSmesh does not require this assumption. The continuation methods[24−26] are of another category. These methods mesh the implicit surface by growing current polygonization's border through the predictor−corrector method, which predicts the next surface point in the tangent direction of the current one and corrects it on the surface. The predictor−corrector method is used in TMSmesh to generate the next corrected point on the surface from current one, and the topology connection is confirmed by checking the continuity between the corrected and the current points, otherwise we restart the predictor−corrector from the current point with a smaller step size, until the continuity is fulfilled. The above process is defined as the trace technique in this paper, and it can be seen as a generalization of the predictor−corrector method. The quality of mesh triangles is well controlled in continuation methods, but techniques for avoiding overlapping, filling the gap between adjacent branches, and selecting proper initial triangles are required. In TMSmesh, no problems of overlapping, gap filling, and selecting initial seeds need to be considered, because the Gaussian surface is polygonized by connecting presampled surface points.

This paper is organized as follows. In Method Section, we present our method for polygonizing the Gaussian surface. Some examples and applications are presented in the Results Section. The final section, Conclusion, gives some concluding remarks.

## Method

In this section, we describe the algorithm for meshing the Gaussian surface. Our algorithm contains two stages. The first stage is to compute the points on the surface by solving a nonlinear equation $\phi(\vec{x}) = t_0$. The second stage is to polygonize the Gaussian surface by connecting the generated points. In the following subsections, each stage is described in detail.

**Computing the Points on the Gaussian Surface.** From the definition of Gaussian surface, the points on the Gaussian surface are the roots of nonlinear equation $\phi(x, y, z) = t_0$, where $\phi(x, y, z)$ is defined in eq 2. Therefore, solving $\phi(x, y, z) = t_0$ is equivalent to computing the points on the Gaussian surface. In this method the equation is solved by the following steps.

Suppose the molecule is placed on a three-dimensional orthogonal grid consisting of $n_x \times n_y \times n_z$ cubes. For an arbitrary cube $[x_i, x_i + h] \times [y_i, y_i + h] \times [z_i, z_i + h]$, where $(x_i, y_i, z_i)$ is the lower-left front corner, and $h$ is the edge length of the cube. To decide whether the cube has an intersection with the surface, we proposed the following lower and upper bounds of $\phi(x)$ in the cube for Gaussian surface:

$$L_i = \sum_{k=1}^{N} e^{-d} L_{k,i}^x L_{k,i}^y L_{k,i}^z \le \phi(x, y, z) \le \sum_{k=1}^{N} e^{-d} U_{k,i}^x U_{k,i}^y U_{k,i}^z = U_i \quad (3)$$

for $(x, y, z) \in [x_i, x_i + h] \times [y_i, y_i + h] \times [z_i, z_i + h]$, where

$$U_{k,i}^{\alpha} = \begin{cases} 1, & c_\alpha^k \in [\alpha_i, \alpha_i + h] \\ \max\{e^{d(\alpha_i - c_\alpha^k)^2/r_k^2}, e^{d(\alpha_i + h - c_\alpha^k)^2/r_k^2}\}, & c_\alpha^k \notin [\alpha_i, \alpha_i + h] \end{cases} \quad (4)$$

$$L_{k,l}^{\alpha} = \min\{e^{d(\alpha_i - c_\alpha^k)^2/r_k^2}, e^{d(\alpha_i + h - c_\alpha^k)^2/r_k^2}\} \quad (5)$$

with $\alpha \in \{x, y, z\}$ and $\vec{c}_k = (c_x^k, c_y^k, c_z^k)$. $U_{k,i}^{\alpha}$ and $L_{k,i}^{\alpha}$ are the upper and lower bounds along $\alpha$-dimension of the kernel located at atom $k$ in the cube, respectively. $U_{k,i}^{\alpha}$ and $L_{k,i}^{\alpha}$ take either 1 or a value of the kernel at the boundary of the cube. If $t_0 \in [L_i, U_i]$, then the Gaussian surface $\phi(x, y, z) = t_0$ may have an intersection with cube $[x_i, x_i + h] \times [y_i, y_i + h] \times [z_i, z_i + h]$, otherwise there is no surface point in the cube. Note that the upper-bound $U_i$ and the lower-bound $L_i$ depend on the edge length of the cube $h$. The bounds are sharper when $h$ is smaller. Above estimation is easy to combine with an octree data structure to decide intersection more adaptively.

The above estimation technique allows the deletion of the majority of cubes, which do not intersect the surface. In each of the left cubes, some surface points are sampled through root finding. Suppose the cube $[x_{i_0}, x_{i_0} + h] \times [y_{i_0}, y_{i_0} + h] \times [z_{i_0}, z_{i_0} + h]$ is one of them, we solve the nonlinear equation $\phi_{ij}(x) \triangleq \phi(x, y_{i_0} + i\tilde{h}, z_{i_0} + j\tilde{h}) = t_0$, for each $\{i, j\}$, $i, j = 1, ..., [h/\tilde{h}]$. The $\tilde{h}$ is to control the vertex density of the mesh. To find the roots, $\phi(x, y_i, z_j)$, $x \in [x_{i_0}, x_{i_0} + h]$, is approximated by the following $M$th-degree polynomial:

$$p_{ij}(2(x - x_{i_0})/h + 1) = \sum_{n=1}^{M} (2i + 1) a_n L_n[2(x - x_{i_0})/h + 1]/2 \quad (6)$$

where $a_n = \int_{-1}^{1} \phi[hx/2 + (x_{i_0} + h/2), y_i, z_j] L_n(x) dx$, $L_n(x)$ is the $n$th-degree Legendre polynomial, $M$ is set as 10, and $h$ is 4 Å in our work. Then $p_{ij}[2(x - x_{i_0})/h + 1] = t_0$ is solved using Jenkins−Traub method.[27] The real roots of $p_{ij}[2(x - x_{i_0})/h + 1] = t_0$ in $[x_{i_0}, x_{i_0} + h]$ should be checked by $|\phi(x, y_i, z_j) - t_0| < \varepsilon$ ($\varepsilon$ is an error tolerance) and be improved by Newton iterations, if needed. This process may lose some roots of $\phi_{ij}(x) = t_0$, due to approximation of $p_{ij}[2(x - x_{i_0})/h + 1]$ to $\phi_{ij}(x)$, but they would be found through trace processes in the polygonization stage.

**Trace Step.** In this subsection, the trace step employed in polygonization stage is described in detail. The objective of the trace step is to connect two (previously identified) grid−surface intersection points. In the trace step, the next connected surface point is predicted and corrected from the initial point with an initial step size, and the connection is confirmed through checking the continuity between the two points. If the continuity is not fulfilled, then restart the prediction and the correction process from the initial point with a smaller step size. Because every trace step is performed either on $xy$- or $yz$-planes between lines parallel to the $x$-axis in the polygonizaton stage, we discuss the details of the trace step on the $xy$-plane between two lines parallel to the $x$-axis as an example. Suppose there are two lines $y = y_0$ and $y = y_1$ on the xy-plane with four surface points $(x_0, y_0)$, $(x_1, y_0)$, $(x_2, y_1)$, $(x_3, y_1)$ on them and $y_1 > y_0$, $\phi_x(x_0, y_0) > 0$, Algorithm 1 connects $(x_0, y_0)$ and $(x_1, y_0)$ (see Figure 1). Moreover, the extreme point $(x^*, y^*)$ is caught during the trace step to preserve the details of the surface.

In the step 2 of Algorithm 1, $\vec{p}_0'$ is the predicted surface point from $\vec{p}_0$ along the tangent direction at $\vec{p}_0$ with step size $h_2$. Step 3 is to correct $\vec{p}_0'$ back to the surface along the gradient direction of $\phi(x)$ at $\vec{p}_0'$. Step 4 is to check whether $\vec{p}_1$ is an extreme point along the $x$-direction. In step 5, condition ($\bigstar_1$) is used to determine if the step-size $h_2$ is acceptable through checking whether the angle between the normal directions at $\vec{p}_0$ and $\vec{p}_1$ is small enough, otherwise restart the prediction and the correction from $\vec{p}_0$ with a

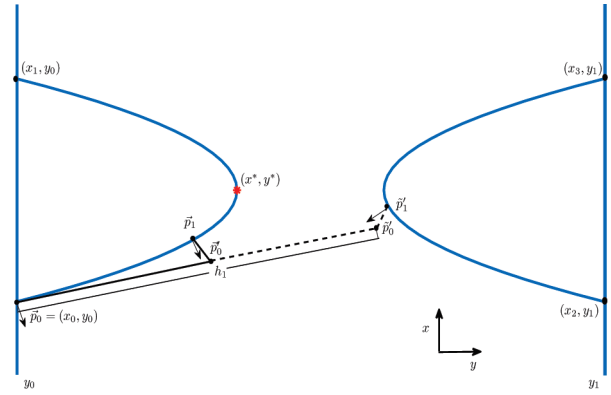



***Figure 1.*** Schematic picture of Algorithm 1. Curved lines indicate the surface on the $xy$-plane. Arrows on the surface denote the normal directions of the Gaussian surface on the $xy$-plane pointing to the outside of the molecule. The $\vec{p}_0$ is the initial point, $(x^*, y^*)$ is an extreme point along $x$-direction, and $(x_1, y_0)$ is the final connected point on the surface obtained through the trace step from $\vec{p}_0$. As an illustration, $\tilde{x}_0'$ and $\tilde{x}_1$ on the dashed line that kinks near the curve on the right-hand side are the predicted and corrected points with a larger initial step $h_1$, which can be avoided through conditions ($\bigstar_1$) in algorithm 1. While $\vec{p}_0'$ is the predicted point along the tangent direction of $\vec{p}_0$ with a smaller step size $h_1/2$, and $\vec{p}_1$ is the corrected surface point from $\vec{p}_0'$.

**Algorithm 1** Trace step

**Input:** Initial step size $h_1$ and initial surface point $\vec{p}_0$. The y-coordinates of two adjacent lines on xy-plane, $y_0$ and $y_1$. User defined small positive value $\varepsilon$ and the bound for the cosine value $\delta$ ($0<\delta<1$).

Step 1, initialize $h_2 = h_1$ and $\vec{p}_0 = (x_0, y_0)$.

Step 2, let $\vec{p}_0' = \vec{p}_0 + h_2(-\phi_y(\vec{p}_0), \phi_x(\vec{p}_0))$.

Step 3, use Newton iterations to find $t$, s.t.

$$\phi(\vec{p}_0' + t(\phi_x(\vec{p}_0'), \phi_y(\vec{p}_0'))) = t_0.$$

Let $\vec{p}_1 = \vec{p}_0' + t(\phi_x(\vec{p}_0), \phi_y(\vec{p}_0))$.

Step 4, if $|\phi_x(\vec{p}_1)| < \varepsilon$, $\vec{p}_1$ is an extreme point along $x$ direction, add it to the extreme point list.

Step 5, if $\cos((\phi_x(\vec{p}_1), \phi_y(\vec{p}_1)), (\phi_x(\vec{p}_0), \phi_y(\vec{p}_0))) < \delta$ (condition $\star_1$) or $(\phi_x(\vec{p}_1)\phi_x(\vec{p}_0) < 0$ and $\min(|\phi_x(\vec{p}_1)|, |\phi_x(\vec{p}_0)|) > \varepsilon)$ (condition $\star_2$), let $h_2 = h_2/2$ and go to step 2.

Step 6, if $(\vec{p}_0(y) - y_0)(\vec{p}_1(y) - y_0) < 0$ (or $(\vec{p}_0(y) - y_1)(\vec{p}_1(y) - y_1) < 0$) [a], interpolate $\vec{p}_0$ and $\vec{p}_1$ to get the connected point $(x_1, y_0)$ (or $(x_1, y_1)$), let $\vec{p}_1 = (x_1, y_0)$ (or $\vec{p}_1 = (x_1, y_1)$) and stop.

Step 7, let $\vec{p}_0 = \vec{p}_1$ and go to step 1.

**Output:** The final connected surface point $(x_1, y_0)$ on $y = y_0$ (or $(x_1, y_1)$ on $y = y_1$), and the extreme point(s) if exist.

[a] Where $\vec{p}_0(y)$ and $\vec{p}_1(y)$ denote y-coordinates of point $\vec{p}_0$ and $\vec{p}_1$, respectively.

smaller step size; $\delta$ is a user-specified bound for cosine value of the angle. If the condition is not sufficient in some cases, then other conditions can be added, such as continuity of higher order derivatives. In the case that an extreme point along the $x$-direction exists between $\vec{p}_0$ and $\vec{p}_1$, condition ($\bigstar_2$) is used to detect it. In step 6, if the line segment connecting $\vec{p}_0$ and $\vec{p}_1$ crosses the line $y = y_0$ (or $y = y_1$), then the point of intersection is the final trace point. In step 7, $\vec{p}_0$ is replaced by $\vec{p}_1$ and starts tracing the next connected surface point from step 1. This process indicates that the final connected point can be located through trace step from initial point, therefore, the position of the final point needs not be known before the trace process. For this reason, a disjointed part of the whole surface will not be missed after the polygonization stage unless no points from the disjointed part are found in the stage of the surface point sampling.

**Polygonization.** This section is devoted to the polygonization step which connects the presampled points obtained through the process described in the section of computing surface points. Because solving $\phi(x, y, z) = t_0$ for different $y$, $z$ values is equivalent to finding the intersection points of the surface and the different lines parallel to $x$-axis, polygonization of the whole surface can be achieved through connecting these points on every adjacent four lines. The problem is how to connect the surface points on the adjacent four lines. Suppose we have surface points set $P$ consisting of four lists of points:

$$\{x_1^i, y_0, z_0\}, \quad i = 1, ..., n_1 \tag{7}$$

$$\{x_2^i, y_0 + \tilde{h}, z_0\}, \quad i = 1, ..., n_2 \tag{8}$$

$$\{x_3^i, y_0, z_0 + \tilde{h}\}, \quad i = 1, ..., n_3 \tag{9}$$

$$\{x_4^i, y_0 + \tilde{h}, z_0 + \tilde{h}\}, \quad i = 1, ..., n_4 \tag{10}$$

in the four adjacent lines:

$$\begin{cases} y = y_0 \\ z = z_0 \end{cases}, \begin{cases} y = y_0 + \tilde{h} \\ z = z_0 \end{cases}, \begin{cases} y = y_0 \\ z = z_0 + \tilde{h} \end{cases}, \begin{cases} y = y_0 + \tilde{h} \\ z = z_0 + \tilde{h} \end{cases} \tag{11}$$

Without loss of generality, assume $n_1 > 0$ and $(x_1^1, y_0, z_0) \triangleq \vec{p}_0$ is chosen to be the initial point. Through invoking the trace step described in former subsection, Algorithm 2 is to connect the points on the four adjacent lines to form small closed loops, i.e., polygons, on the surface.

Algorithm 2 is repeated until all points are connected, i.e., $P$ is empty. This algorithm traces vertices of polygons in $xy$- and $xz$-planes alternately. The variable idx records the location of the last trace step, and the $\text{sig}_x$ records the direction of the first trace step. If the traced vertex is not in the same $xy$- or $xz$-plane as $\vec{p}_0$, then the next trace direction will be reversed. After Algorithm 2 is finished, $\vec{p}_j$, $j = 0, ..., i - 1$, and the extreme points (if they exist) along the $x$-direction obtained during the trace steps are connected and form a polygon on the surface. Figure 2 shows some

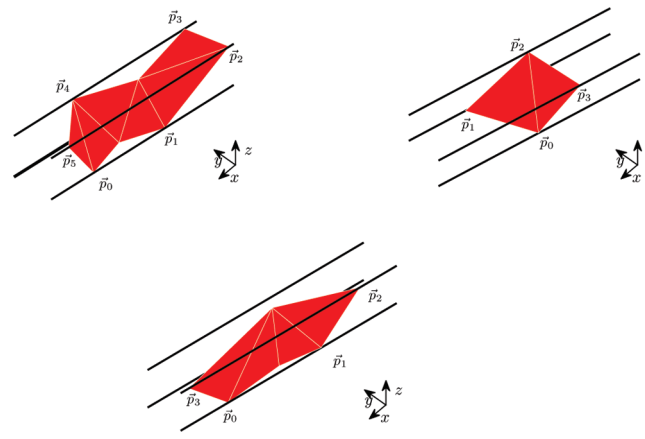



***Figure 2.*** Some polygons with different numbers of vertices obtained from Algorithm 2. The unlabeled vertices are extreme points obtained from the trace steps.

**Algorithm 2** Connecting the points on the four adjacent lines from the initial point $\vec{p}_0$ to form one polygon.

**Input**: The set $P$ containing coordinates of all sampled points on the four adjacent lines. The grid space $\tilde{h}$. The yz-coordinates of lower-left line $\{y_0, z_0\}$.The initial point $\vec{p}_0$.

Find the connected point $\vec{p}_1$ on adjacent line using trace step on xy-plane from the initial point $\vec{p}_0$ along the tangent direction $(-sig_x\phi_y(\vec{p}_0), sig_x\phi_x(\vec{p}_0), 0)$, where $sig_x$ is the sign of $\phi_x(\vec{p}_0)$.
Let $idx = 1$, $i = 1$ and $P = P - \{\vec{p}_0\}$.
**while** $\vec{p}_i \neq \vec{p}_0$ **do**
  **if** $\vec{p}_i$ is in set $P$ **then**
    Let $P = P - \{\vec{p}_i\}$.
  **end if**
  **if** $idx = 1$ **then**
    **if** $\vec{p}_i(y) = y_0$ **then**
      Find the connected point $\vec{p}_{i+1}$ on adjacent line using trace step on xz-plane from $\vec{p}_i$ along the tangent direction $(-sig_x\phi_z(\vec{p}_i), 0, sig_x\phi_x(\vec{p}_i))$.
    **else if** $\vec{p}_i(y) = y_0 + \tilde{h}$ **then**
      Find the connected point $\vec{p}_{i+1}$ on adjacent line using trace step on xz-plane from $\vec{p}_i$ along the tangent direction $(sig_x\phi_z(\vec{p}_i), 0, -sig_x\phi_x(\vec{p}_i))$.
    **end if**
    Let $idx = 2$.
  **else if** $idx = 2$ **then**
    **if** $\vec{p}_i(3) = z_0$ **then**
      Find the connected point $\vec{p}_{i+1}$ on adjacent line using traces step on xy-plane from $\vec{p}_i$ along the tangent direction $(-sig_x\phi_y(\vec{p}_i), sig_x\phi_x(\vec{p}_i), 0)$.
    **else if** $\vec{p}_i(3) = z_0 + \tilde{h}$ **then**
      Find the connected point $\vec{p}_{i+1}$ on adjacent line using trace step on xy-plane from $\vec{p}_i$ along the tangent direction $(sig_x\phi_z(\vec{p}_i), -sig_x\phi_x(\vec{p}_i), 0)$.
    **end if**
    Let $idx = 1$.
  **end if**
  Let $i = i + 1$.
**end while**
**Output**: The polygon whose vertices are $\vec{p}_j, j = 1, \cdots, i$, and the extreme points (if they exist) along x-direction obtained during the trace steps.

examples of polygons with a different number of vertices obtained with Algorithm 2. The polygon can be simpler when the distances between the adjacent lines are shorter due to the smoothness of Gaussian surface. Based on the polygonized surface, the triangulation of the surface can be produced using standard polygon triangulation methods.[28]

## Results

**Performance.** In this section, the performance of TMSmesh is compared with those of LSMS and MSMS. LSMS is a very fast program using a level-set method to present the surface based on cubic grids. MSMS is a typical and efficient software to triangulate the SES in the modeling area. A set of biomolecules with different sizes is chosen as a test benchmark. The meshing softwares are run on the molecular PQR files (PDB + atomic charges and radii information). For tests of the mesh tool and its applications in this work, we prepare a PQR benchmark (see Table 1) that can be found and is downloadable at our web page http://lsec.cc.ac.cn/~lubz/Meshing.html. It is worth making a note here about the vertex density used in TMSmesh and LSMS for comparison with MSMS surface density. For TMSmesh, grid spaces of 1.0 and 0.7 Å are chosen to approximate the molecular surface vertex densities 1/Å$^2$ and 2/Å$^2$, respectively. For LSMS, the current implementation works only on the following grid sizes: $16^3$, $32^3$, $64^3$, $128^3$, $256^3$, and $512^3$ (requiring a 4GB memory machine). Therefore, for each molecule, a proper grid size in LSMS is chosen to achieve the approximate density of 1/Å$^2$ or 2/Å$^2$, according to the maximum molecular length in *xyz* directions.

Table 2 shows the CPU time and memory use for these methods with 1 and 2 vertex/Å$^2$ mesh density. All computations run on Dell Precision T7500 with Intel(R) Xeon(R) CPU 3.3 GHz and 48GB memory under 64bit Linux system. As shown in Table 2, TMSmesh costs less memory than LSMS and MSMS but much more CPU time for small- or medium-sized molecules. The main cost of TMSmesh is in the polygonization stage that connects the presampled surface points on parallel lines through invoking the trace steps intensively. During each trace step, prediction and correction need to be performed several times with a small step size about 0.1 to 0.2 Å, i.e., there needs to be 5−10 prediction−correction steps within 1 Å distance on the surface to ensure the continuity of curves connecting vertices. However, LSMS directly searches and approximates the molecular surface based on cubic grid points using the level-set method. MSMS analytically computes molecular surface by first generating

***Table 1.*** Description of Molecules in the PQR Benchmark

| molecule (name or PDB code) | number of atoms | description |
|---|---|---|
| GLY | 7 | a single glycine residue |
| ADP | 39 | ADP molecule |
| 2JM0 | 589 | PDB code, chicken villin headpiece subdomain containing a fluorinated side chain in the core |
| FAS2 | 906 | fasciculin2, a peptidic inhibitor of AChE |
| AChE monomer | 8280 | mouse acetylcholinesterase monomer |
| AChE tetramer | 36 638 | the structure of AChE tetramer, taken from ref 29 |
| 30S ribosome | 88 431 | 30S ribosome, the PDB code is 1FJF |
| 70S ribosome | 165 337 | obtained from 70S_ribosome3.7A_model140.pdb.gz on http://rna.ucsc.edu/rnacenter/ribosome_downloads.html |
| 3K1Q | 203 135 | PDB code, a backbone model of an aquareovirus virion |
| 2X9XX | 510 727 | a complex structure of the 70S ribosome bound to release factor 2 and a substrate analog, which has 4 split PDB entries: 2X9R, 2X9S, 2X9T, and 2X9U[30] |
| 1K4R | 1 082 160 | PDB code, the envelope protein of the dengue virus[31] |

***Table 2.*** CPU Time and Memory Use for Molecular Surface Generation by TMSmesh, LSMS, and MSMS[a]

| | | CPU time (S) | | | memory use (MB) | | |
|---|---|---|---|---|---|---|---|
| molecule | number of atoms | TMSmesh | LSMS[b] | MSMS | TMSmesh | LSMS | MSMS |
| FAS2 | 906 | 1.2 | 0.05 | 0.1 | 2 | 10 | 2 |
| | | 1.8 | 0.1 | 0.1 | 4 | 19 | 2 |
| AChE monomer | 8280 | 12 | 0.1 | 0.6 | 11 | 21 | 21 |
| | | 19 | 0.4 | 0.8 | 19 | 33 | 21 |
| AChE tetramer | 36 638 | 40 | 0.5 | 5.9 | 36 | 346 | 75 |
| | | 62 | 3.6 | 7.1 | 54 | 257 | 79 |
| 30S ribosome | 88 431 | 92 | 3.6 | 16.2 | 64 | 260 | 198 |
| | | 151 | 28.1 | 19.1 | 100 | 2016 | 212 |
| 70S ribosome | 165 337 | 180 | 3.8 | 46.2 | 127 | 262 | 469 |
| | | 283 | 28.6 | Fail | 185 | 2100 | – |
| 3K1Q | 203 135 | 226 | 4.0 | 51.5 | 131 | 262 | 383 |
| | | 359 | 28.9 | 55.1 | 192 | 2100 | 410 |
| 2X9XX | 510 727 | 577 | 30.5 | Fail | 271 | 2100 | – |
| | | 910 | Fail | Fail | 410 | – | – |
| 1K4R | 1 082 160 | 1260 | 30.5 | Fail | 630 | 2168 | – |
| | | 2080 | Fail | Fail | 890 | – | – |

[a] The data in the first and second row for each molecule are corresponding to the density 1 vertex/Å$^2$ and 2 vertex/Å$^2$, respectively. [b] The failed cases in this column require a grid size of $1024^3$ or larger, which is not supported by LSMS.

the so-called reduced surface that is obtained directly from atomic geometry information. Both LSMS and MSMS avoid the time-consuming step, polygonization. This makes LSMS and MSMS cost less CPU time than that of TMSmesh. Nevertheless, either the surface topology or the smoothness may not be guaranteed in LSMS and MSMS. TMSmesh is expected to be speeded up using an adaptive box structure, parallel computing, and more sophisticated polygonization algorithm. For LSMS, the cost is proportional to $L^3$, where $L$ is the number of grids in one dimension. Therefore, the memory requirement and the CPU time increase dramatically when $L$ becomes large. For MSMS, the computational complexity is $O[N\,log(N)]$, where $N$ is the number of atoms, but the singularity of the molecular surface may cause numerical instability and produce incorrect results. In TMSmesh, the number of cubes is proportional to the number of atoms, since the edge length of cube is fixed to be 4 Å in our work. In addition, the calculations are done locally, and no global information is needed during the process of estimating the bounds of $\phi(x)$ in each cube, computing surface points, and tracing of the left cubes intersecting the surface. The reason is that calculating the values of $\phi(x)$ and its gradients only need to sum Gaussian kernels for near

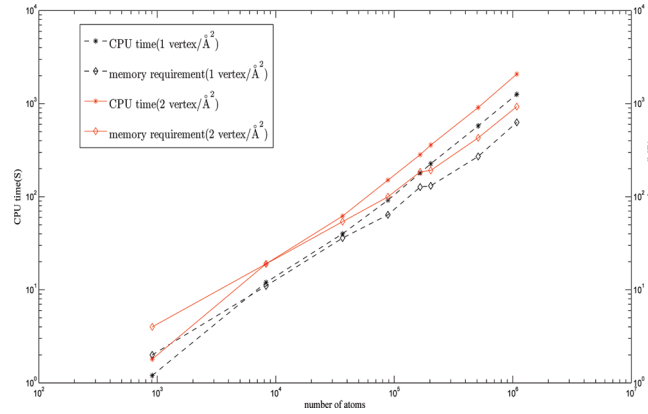



***Figure 3.*** Computational performance of TMSmesh.

atoms, as the Gaussian kernel $e^{d(\|\vec{x}-\vec{c}_i\|^2/r_i^{2-1})}$ decreases to 0 faster when $\|\vec{x}-\vec{c}_i\|$ is large. As shown in Table 2 and Figure 3, the complexity of TMSmesh is $O(N)$. Compared to LSMS and MSMS, TMSmesh produces triangulations of a smooth surface, and it can be successfully applied to a biomolecule consisting of more than one million atoms, such as the dengue virus as shown in Figure 4. Because the virus structure is among the largest ones in the Protein Data Bank

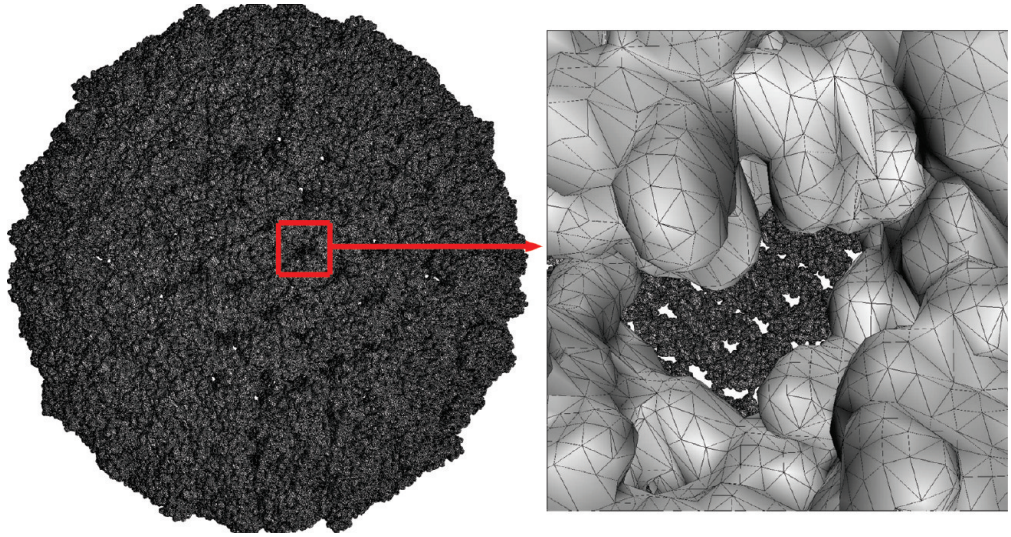


***Figure 4.*** Surface triangular mesh of the envelope protein of the dengue virus (PDB code 1K4R).[31] The left-hand side is the whole surface mesh, and the right-hand side is a close view of a selected part with a gap on the surface (surrounded by the box). Because the structure is a shell, the inner surface of the other side of the shell is also shown through the gap. The mesh density is 1 vertex/Å$^2$.

(PDB), together with consideration of good algorithm stability, TMSmesh can be expected to be capable of handling the arbitrary size of molecules available in PDB.

Because MSMS is a widely used tool for surface meshing in molecular modeling, we compare the qualities, in particular uniformness, of triangles produced by TMSmesh and MSMS. The distributions of the ratios of the shortest to longest edge lengths of each triangle are used to describe the uniformness of meshes. A ratio of 1.0 corresponds to an equilateral triangle, and a ratio close to 0 indicates poor uniformness of the triangle. In other words, the higher the ratio, the better the quality of the triangle. The ratio distributions of meshes for a large molecule, AChE tetramer, and a relatively small one, FAS2, are shown in Figure 5. TMSmesh and MSMS meshes with two densities 2 and 10 vertex/Å$^2$ are compared. It is shown that at both low and high densities, the ratios of TMSmesh meshes are clustered around 0.75. Comparatively, at a low density of 2 vertex/Å$^2$, the ratios of MSMS meshes distribute more evenly in [0,1] (see the first and third rows of Figure 5). At a high density of 10 vertex/Å$^2$ (see the second and the bottom rows of Figure 5), the distributions of MSMS mesh ratios are improved and clustered around 0.65 that still indicates less uniform triangle clusters than TMSmesh results; furthermore, at the region with a ratio close to 0, TMSmesh also generates less percentage of such poor quality triangles than MSMS. In addition, successful applications (see the following subsection) of all of our meshes to boundary element simulations also indicate improvement in mesh quality relative to MSMS mesh in the sense of right topology (e.g., without single-element-connected edges, isolated points), uniformness, and smoothness.

Finally, it is worth making a note of the molecular cavity as explored by many other tools. TMSmesh does not differentiate the outer and interior surfaces of cavities in the meshing process. The cavities can be located through the connectivities of the triangle elements, because an internal cavity is a disjointed component of Gaussian surface (eq 1), and its normal directions $\nabla\phi(x)$ are inward. The same method of finding internal cavities is used in GRASP.[12]

**Application to Boundary Element Simulation of Electrostatics.** Similar to other surface generation softwares, such as MSMS and LSMS, the surface mesh generated by TMSmesh preserves molecular surface features and thus can be applied to molecular visualization and analysis of surface area, topology, and volume in computational structure biology and structural bioinformatics. Furthermore, the goal of this work is to extend applications to some advanced mathematical modeling of biomolecules, which places demands upon the quality and the rigorous topology of the meshes.

In this work, we test the meshes in the usage of a boundary element method to calculate the Poisson−Boltzmann electrostatics. The BEM software used is a publicly available PB solver AFMPB.[32] MSMS meshes have already been used in many previous boundary element PB works for smaller molecules and have demonstrated to generate reasonable results. Because LSMS mesh is built on cubic grid that can deviate somehow from the curved molecular surface (unless the grid space is small enough), we did not perform AFMPB computation with LSMS mesh. Instead, the AFMPB results from meshes of TMSmesh and MSMS are compared. Our test cases, up to molecular size of 2X9XX (which is a complex structure of the 70S ribosome bound to release factor 2 and a substrate analog)[30] due to the memory limit of our machine, show that AFMPB can go through and produce converged results with all of our meshes for the molecules described in Table 2. However, the solver fails for meshes directly obtained with MSMS for 30S ribosome and larger molecules in Table 2, which is due to singularities or incorrect topologies in MSMS meshes. Figure 6 shows the solvation energies by AFMPB as well as the surface areas and molecular volumes computed from the meshes of three small molecules, GLY, ADP, and 2JM0 (see Table 1) using different mesh densities. The figure indicates that BEM

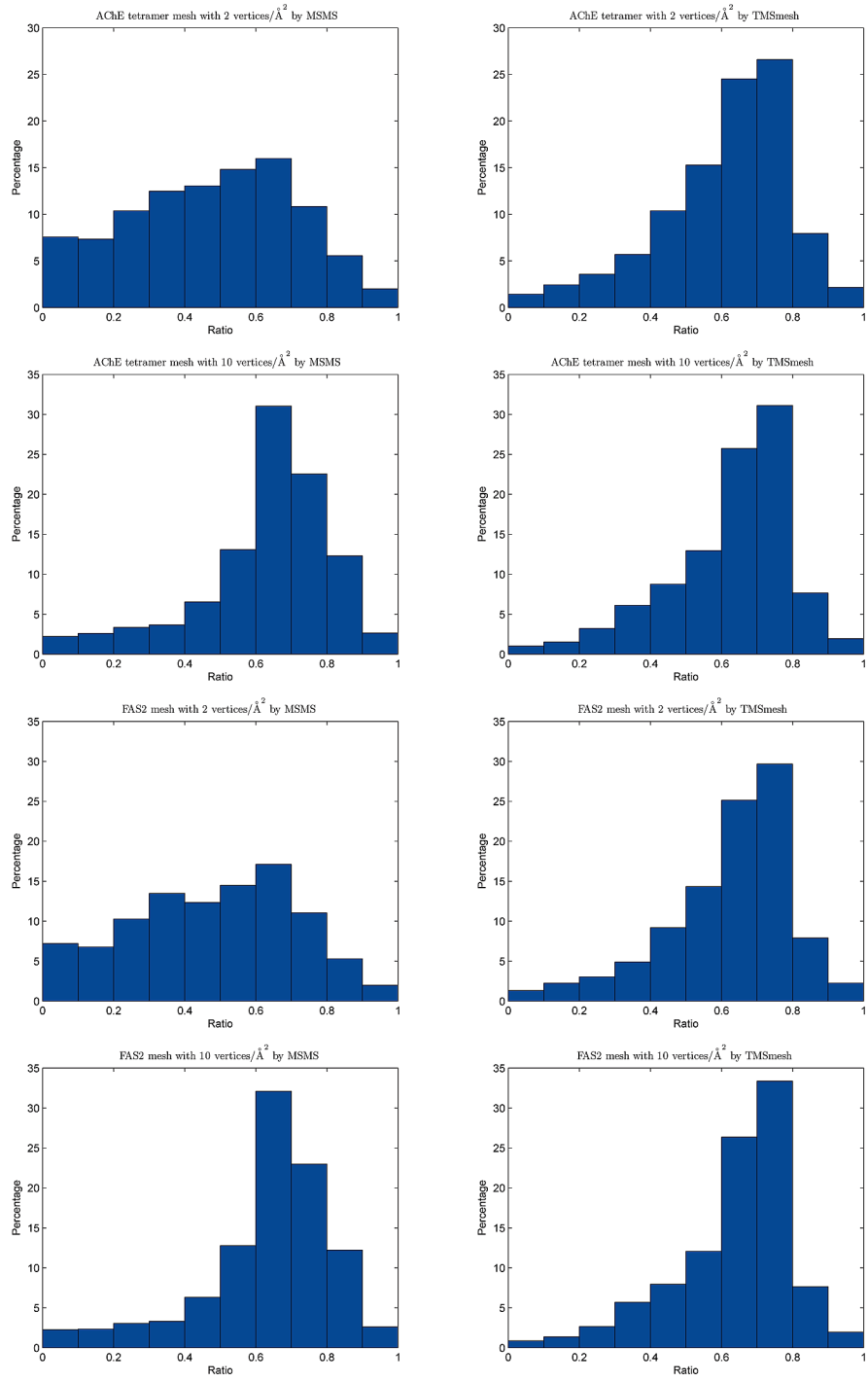



***Figure 5.*** Distributions of ratio of the shortest edge length to the longest edge length of each triangle produced by MSMS (left column) and TMSmesh (right column). The first two rows are for AChE tetramer meshes with densities of 2 and 10 vertex/Å$^2$, respectively. The last two rows are for FAS2 meshes with densities of 2 and 10 vertex/Å$^2$, respectively. A ratio of 1.0 corresponds to an equilateral triangle.

calculations work for all the meshes with different densities produced by TMSmesh, and the mesh leads to convergent and reasonable results for energy, area, and volume when the mesh resolution is increased. This is reasonable because the mesh converges to the implicitly defined Gaussian surface with the increasing of the resolution. Figure 6 also shows that the results computed by TMSmesh converge more smoothly than those of MSMS as the number of triangles increased. There are some discrepancies between the quantities calculated with TMSmesh and MSMS meshes, which is due to the different definitions of molecular surface used in the mesh tools.

It is known that BEM is a memory-saving approach to solve the PBE for macromolecules, which indicates that a combination of TMSmesh and a BEM solver, such as AFMPB, is a promising way to handle large-scale molecular systems for electrostatic calculations. As a representative molecular system, we choose the complex structure 2X9XX (see Table 1), which contains 510 727 atoms with a dimension of 270 × 392 × 384 Å (Figure 7). The molecular surface

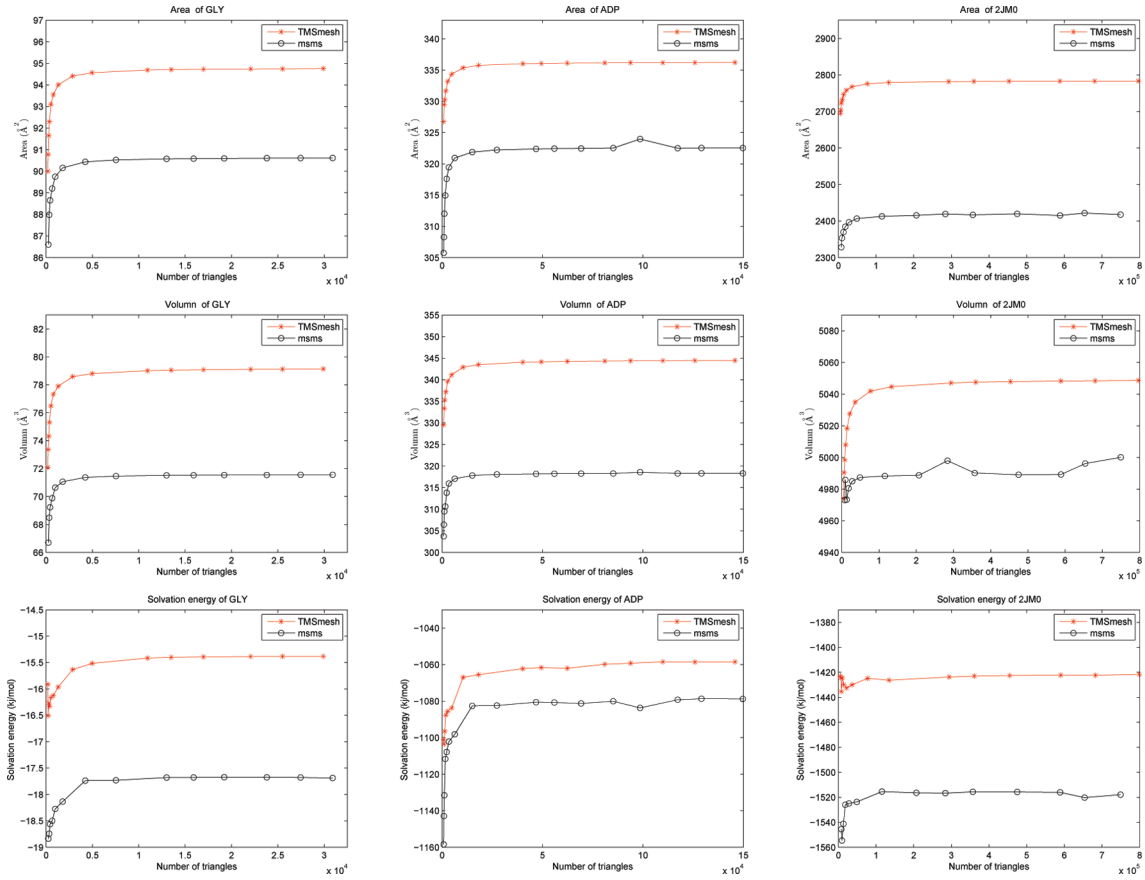



***Figure 6.*** Area (first row), volume (second row), and solvation energy (third row) for GLY (left column), ADP (middle column), and 2JM0 (right column).

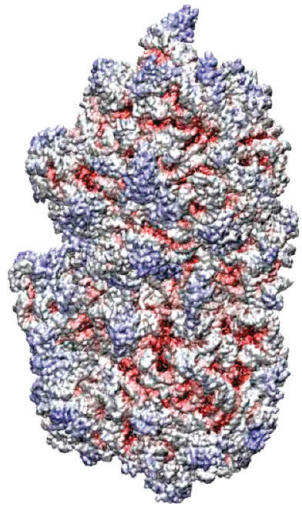


***Figure 7.*** Electrostatic potential surface of the complex structure 2X9XX calculated with AFMPB[32] using a surface mesh density 1 vertex/Å$^2$.

discretization using TMSmesh with a mesh density of 1 vertex/Å$^2$ results in 3 693 500 triangular elements and 1 841 858 nodes. Figure 7 shows the computed electrostatic potentials mapped on the molecular surface.

## Conclusion

We have described a new method for molecular surface meshing by a tracing surface technique. The implemented software TMSmesh is shown as a robust tool for meshing molecular Gaussian surfaces in the sense that: (1) It can stably handle arbitrary sizes of molecules available in PDB on a typical desktop or laptop machine, even for the not "good" molecular structures (such as ones with strong atomic clash) and (2) the generated mesh has good quality (smoothness, uniformness, and topological correctness). The mesh converges to the smooth Gaussian surface when the mesh resolution increased and from which all the calculations of electrostatic solvation energy, surface area, and molecular volume show good convergence performance and reasonable results. Therefore, in addition to usual applications of molecular visualization and geometry analysis, the mesh is also shown to be applicable for numerical simulations with boundary element methods. Specifically, a combination of TMSmesh and BEM solver opens a possible route to simulate electrostatic solvation of large-scale molecular systems on a desktop computer.

In order to simulate more complicated and wider ranges of biophysical processes using a variety of numerical techniques and modeling approaches, the current meshing method needs further improvements. First, efficiency seems to be the current bottleneck in some possible applications where the mesh needs to be either generated for large systems or generated frequently, such as in multiple-conformational analysis, BEM-based implicit solvent MD simulations[33,34] (whereas in some finite difference-based MD simulations,[35] surface meshing is not required), or elastic modeling of

conformational changes. Second, surface mesh smoothness needs to be further improved. Third, molecular volume mesh generation based on surface mesh is required for some finite element types of simulations. In addition, in the case of higher order numerical computations, the Gaussian surface may need to be discretized more accurately using curved elements. It would be convenient to generate curved elements by small modification of TMSmesh, because the method produces an abundance of trace information in the meshing procedure.

Finally, it is worth a mention regarding PB calculations. It is hard to conclude so far which surface specification is the best for biophysical studies due to being complicated by some other factors (like the atomic radii) in the setup of a PB calculation that can also affect the final results. Likewise, the Gaussian surface model and the meshing approach adopted in this work for PB electrostatic calculations will need further systematic studies and comparisons with experiments or other computational methods.

**Acknowledgment.** We thank the reviewers for their comments and suggestions, which are very helpful for improving the paper. M.X. was supported in part by the China NSF (NSFC20872107). B.Z. was supported by the Chinese Academy of Sciences, the State Key Laboratory of Scientific/Engineering Computing, and the China NSF (NSFC10971218).

**Note Added after ASAP Publication.** This article was published on the web on November 30, 2010. Algorithm 1 and Algorithm 2 have been revised to remove the color. The correct version was published on December 3, 2010.

## References

(1) Lu, B. Z.; Zhou, Y. C.; Holst, M. J.; McCammon, J. A. *Commun. Comput. Phys.* **2008**, *3*, 973–1009.

(2) Edelsbrunner, H. *Discrete Comput. Geom.* **1999**, *21*, 87–115.

(3) Bates, P. W.; Wei, G. W.; Zhao, S. *J. Comput. Chem.* **2008**, *29*, 380–391.

(4) Duncan, B. S.; Olson, A. J. *Biopolymers* **1993**, *33*, 231–238.

(5) Blinn, J. F. *ACM Trans. Graph.* **1982**, *1*, 235–256.

(6) Mcgann, M. R.; Almond, H. R.; Nicholls, A.; Grant, A. J.; Brown, F. K. *Biopolymers* **2003**, *68*, 76–90.

(7) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(8) J.Weiser; Shenkin, P.; Still, W. *J. Comput. Chem.* **1999**, *20*, 688–703.

(9) Yun, Z.; Matthew, P.; Richard, A. *J. Comput. Chem.* **2005**, *27*, 72–89.

(10) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.

(11) Connolly, M. L. *Science* **1983**, *221*, 709–713.

(12) Nicholls, A.; Bharadwaj, R.; Honig, B. *Biophys. J.* **1995**, *64*, 166–167.

(13) Vorobjev, Y. N.; Hermans, J. *Biophys. J.* **1997**, 722–732.

(14) Edelsbrunner, H.; Mücke, E. P. *ACM Trans. Graph.* **1994**, *13*, 43–72.

(15) Sanner, M.; Olson, A.; Spehner, J. *Biopolymers* **1996**, *38*, 305–320.

(16) Ryu, J.; Park, R.; Kim, D.-S. *Comput. Aided Des.* **2007**, *39*, 1042–1057.

(17) Zhang, Y.; Ch, G. X.; Bajaj, R. *Comput. Aided Geomet. Des.* **2006**, *23*, 510–530.

(18) Yu, Z.; Holst, M. J.; Andrew McCammon, J. *Finite Elem. Anal. Des.* **2008**, *44*, 715–723.

(19) Can, T.; Chen, C.-I.; Wang, Y.-F. *J. Mol. Graphics Modell.* **2006**, *25*, 442–454.

(20) Chavent, M.; Levy, B.; Maigret, B. *J. Mol. Graphics Modell.* **2008**, *27*, 209–216.

(21) Cheng, H.; Shi, X. *Comput. Geom.* **2009**, *42*, 196–206.

(22) Lorensen, W.; Cline, H. E. *Comput. Graph.* **1987**, *21*, 163–169.

(23) Ju, T.; Losasso, F.; Schaefer, S.; Warren, J. D. *ACM Trans. Graph.* **2002**, *21*, 339−346.

(24) Hilton, A.; Stoddart, A. J.; Illingworth, J.; Windeatt, T. *IEEE Int. Conf. Image Process.* **1996**, 381–384.

(25) Hartmann, E. *Vis. Comput.* **1998**, 95–108.

(26) Karkanis, T.; Stewart, J. *IEEE Comput. Graphics Appl.* **2001**, 60–69.

(27) Jenkins, M.; Traub, J. F. *Numer. Math.* **1970**, 253–263.

(28) de Berg, M.; van Kreveld, M.; Overmars, M.; Schwarzkopf, O. *Computational Geometry: Algorithms and Applications*, 2nd ed.; Springer-Verlag: New York, 2000; pp 45−61.

(29) Zhang, D.; McCammon, J. A. *PLoS Comput. Biol.* **2005**, 484–491.

(30) Jin, H.; Kelley, A. C.; Loakes, D.; Ramakrishnan, V. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, 8593–8598.

(31) Kuhn, R. J.; Zhang, W.; Rossmann, M. G.; Sergei V. Pletnev, J. C.; Lenches, E.; Jones, C. T.; Mukhopadhyay, S.; Paul R. Chipman, E. G. S.; Baker, T. S.; Strauss, J. H. *Cell* **2002**, 715–725.

(32) Lu, B. Z.; Cheng, X. L.; Huang, J. F.; McCammon, J. A. *Comput. Phys. Commun.* **2010**, 1150–1160.

(33) Wang, C. X.; Wan, S. Z.; Xiang, Z. X.; Shi, Y. Y. *J. Phys. Chem. B* **1997**, *101*, 230–235.

(34) Lu, B. Z.; Wang, C. X.; Chen, W. Z.; Wan, S. Z.; Shi, Y. Y. *J. Phys. Chem. B* **2000**, *104*, 6877–6883.

(35) Wang, J.; Tan, C. H.; Tan, Y. H.; Lu, Q.; Luo, R. *Commun. Comput. Phys.* **2008**, *3*, 1010–1031.

CT100376G

JCTC Journal of Chemical Theory and Computation

# The Importance of Going beyond Coulombic Potential in Embedding Calculations for Molecular Properties: The Case of Iso-G for Biliverdin in Protein-Like Environment

Georgios Fradelos and Tomasz A. Wesolowski*

*Université de Genève, Département de Chimie Physique 30, quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland*

Received July 23, 2010

**Abstract:** The importance of the nonelectrostatic component of the embedding potential is investigated by comparing the complexation induced shifts of the iso-g obtained in embedding calculations to its supermolecular counterparts. The analyses are made in view of such multilevel simulations, for which supermolecular strategy is either impractical or impossible, such as the planned simulations for the whole enzyme ferredoxin oxidoreductase. For the biliverdin radical surrounded by a few amino acids, it is shown that the embedding potential comprising only Coulomb terms fails to reproduce even qualitatively the shifts evaluated from supermolecular calculations. The nonelectrostatic component of the exact embedding potential is a bifunctional of two electron densities [Wesolowski and Warshel, *J. Phys. Chem.* **1993**, *97*, 8050; Wesolowski, *Phys. Rev. A* **2008**, *77*, 012504]. Therefore we analyze in detail both the quality of the approximant for the bifunctional and the importance of the choice of the electron densities at which it is evaluated in practical calculations.

## Introduction

Multilevel techniques in numerical simulation apply the embedding potential to couple the subsystem described at the quantum mechanical level with its environment described using simpler descriptors. If the quantum mechanical treatment of the whole system under investigation is impractical or impossible, then such techniques are the only options available. Such approaches are frequently referred as quantum mechanical/molecular mechanical (QM/MM)[1] and are widely applied in simulation of biomolecules,[2,3] materials,[4] liquids,[5] solids,[6,7] and interfaces,[8] for instance. The embedding potential, i.e., the potential added to the potential in the environment-free case, $v_0(\vec{r})$, is commonly represented by the electrostatic potential (atomic units are applied in all the equations given in the present work):

$$v^{\mathrm{eff}}_{\mathrm{emb(Coulomb)}}[\rho_A, \rho_B; \vec{r}] = v^{B}_{\mathrm{ext}}(\vec{r}) + \int \frac{\rho_B(\vec{r}')}{|\vec{r}' - \vec{r}|} \mathrm{d}\vec{r}' \quad (1)$$

The convention used in eq 1 indicating that the embedding potential is a functional of the two electron densities $\rho_A$ (the electron density of the embedded system) and $\rho_B$ (the electron density of the environment) and is used for the sake of the subsequent discussions. Note that the embedding potential given in eq 1 does not depend on $\rho_A$, i.e., the electron density of the embedded system. In practice, the second term representing the classical Coulomb electron−electron repulsion is evaluated using truncated polycenter multipole expansion. Especially simple is the case of the expansion truncated to monopoles, which leads to a very efficient computational method where the nuclear and monopole expansion charges can be bunched together giving rise to a set of distributed effective charges.

The potential given in eq 1 does not take into account the nonelectrostatic effects on the electronic structure of the embedded species which arise from the Pauli exclusion principle (see ref 9 for discussion of this issue in a model system). If such effects cannot be neglected, then the straightforward solution is to add to the electrostatic potential a nonlocal component consisting of projection operators (or simplified molecular pseudopotentials),[10] which hinges on

* Corresponding author. E-mail: tomasz.wesolowski@unige.ch. Telephone: 041223796101.

10.1021/ct100415h 
Published on Web 12/07/2010

the availability of the orbitals for the environment. In multilevel simulations, such as QM/MM, etc., especially if the target is not a property directly related to the electronic structure but to the potential energy surface, the nonelectrostatic component of the exact embedding potential is frequently neglected, resulting in an inexpensive computational approach. In such a case, the deficiencies of electrostatic-only embedding potentials can be corrected by special terms added to the final expression for the total energy.[11−15] Such corrections are, however, strongly system and method dependent as pointed out by several authors. For instance, the authors of a recent review state that, "QM/MM calculations are not yet black-box procedures",[16] and the authors of ref 17 conclude that "...the need to partition the complete system into regions treated at different levels of theory, and to allow these regions to interact, creates two significant problems. First is the treatment of the QM/MM boundary, as it is often necessary to place this boundary across covalent bonds. Second is the parameterization of the hybrid system Hamiltonian." Several authors reported problems, such as excess polarization of the embedded subsystem by the environment, when the embedding potential comprises only electrostatic terms.[18−20] Specific numerical problems arising from neglecting nonelectrostatic terms in the embedding potential, especially if the basis set used to construct embedded orbitals extends into the environment, were also reported.[21−23] In numerical simulations aiming at obtaining properties of embedded systems, which are other than ground-state energy, the large basis sets are frequently indispensable. Therefore, the quality of the calculated values depends directly on the accuracy of the used embedding potential. The pragmatic solution made for the ground-state energy, i.e., as adding a correction term directly to the energy, is less straightforward for other observables. To correct errors for such quantities, the cure must be applied to the embedding potential, as it determines the quality of the calculated electronic structure.

The frozen-density embedding theory (FDET)[24−28] provides a formal basis for computational methods (besides our own work see also other representative papers)[29−33] in which both the energy as well as the electronic properties are evaluated in a self-consistent manner. Below, we outline the basic elements of the frozen-density embedding theory:

- **Basic variables.** The total investigated system is characterized by two quantities: the density $\rho_B(\vec{r})$, which for a given electronic problem is a frozen function, and the density $\rho_A(\vec{r})$, which is represented using auxiliary quantities, such as occupied orbitals of noninteracting reference system $\{\phi_A^i(\vec{r})\}$,[24] occupied and unoccupied orbitals of noninteracting reference system,[25] interacting wave function,[27] or one-particle density matrix.[28]
- **Constrained search.** The density $\rho_A(\vec{r})$ is obtained by performing the following search:

$$\begin{aligned} E_{\mathrm{emb}}[\rho_B] &= \min_{\rho_A \geq 0} E^{HK}[\rho_A + \rho_B] \text{ for } \int \rho_A(\vec{r})\mathrm{d}\vec{r} = N_A \\ &= \min_{\rho \geq \rho_B \geq 0} E^{HK}[\rho] \text{ for } \int \rho_B(\vec{r})\mathrm{d}\vec{r} = N_B \end{aligned} \quad (2)$$

- **Performing the constrained search by modifying the external potential.** The search is conducted in practice by solving the following equation:

$$(\hat{H}_o + \hat{v}_{\mathrm{emb}})\psi = E_{\mathrm{emb}}\psi \quad (3)$$

where $\hat{H}_o$ is the environment-free Hamiltonian, and the $\hat{v}_{\mathrm{emb}}(\vec{r})$ has the form of a local potential ($v_{\mathrm{emb}}^{\mathrm{eff}}(\vec{r})$), which is determined by the pair of densities $\rho_A(\vec{r})$ and $\rho_B(\vec{r})$, hence orbital-free embedding.

- **Orbital-free embedding potential.** The form of the dependence of the embedding potential on the densities $\rho_A(\vec{r})$ and $\rho_B(\vec{r})$ depends on what QM descriptor is used as the auxiliary quantity for $\rho_A(\vec{r})$ and is given in respective publications.[24,27,28] For the following descriptors: orbitals of noninteracting reference system, a wave function of the full configuration interaction form, and one-particle density matrix, the orbital-free embedding potential reads:

$$\begin{aligned} v_{\mathrm{emb}}^{\mathrm{eff}}[\rho_A, \rho_B; \vec{r}] = v_{\mathrm{ext}}^{B}(\vec{r}) + \int \frac{\rho_B(\vec{r}')}{|\vec{r}' - \vec{r}|}\mathrm{d}\vec{r}' + \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho}\right|_{\rho=\rho_A+\rho_B} - \\ \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho}\right|_{\rho=\rho_A} + \left.\frac{\delta T_s[\rho]}{\delta \rho}\right|_{\rho=\rho_A+\rho_B} - \left.\frac{\delta T_s[\rho]}{\delta \rho}\right|_{\rho=\rho_A} \end{aligned} \quad (4)$$

The correspondence given in eq 4 involves density functionals known in the Kohn−Sham formulation[34] of density functional theory,[35] the functional of the exchange−correlation energy ($E_{\mathrm{xc}}[\rho]$), and the functional of the kinetic energy in a noninteracting system ($T_s[\rho]$). The pair of functional derivatives of the functional $T_s[\rho]$ arises from from nonadditivity of this functional and represents a potential denoted as $v_t^{\mathrm{nad}}[\rho_A, \rho_B](\vec{r})$ in the present work.

In this context, it is useful to relate the frozen-density embedding theory to the subsystem formulation of density functional theory (SDFT)[36,37] and to the recently developed partition density functional theory (PDFT).[38] Both SDFT and PDFT lead to the exact ground-state electron density and the energy of the whole investigated system an alternative way to the conventional Kohn−Sham framework. In SDFT, the charges of each subsystem are assumed to be integral (similarly as in FDET), whereas fractional charges of subsystems are allowed in PDFT. The FDET targets not the ground-state electron density of the total system but the density minimizing the Hohenberg−Kohn energy functional for the total system with the presence of constraints. FDET, therefore, can lead to the same total ground-state density as SDFT and Kohn−Sham DFT or PDFT, only when for particular constraints[26] (see also below). In the case of two subsystems, SDFT is based on the following variational principle:

$$E_o = \min_{\rho_A \geq 0, \rho_B \geq 0} E^{HK}[\rho_A + \rho_B] \text{ for } \int \rho_A(\vec{r})\mathrm{d}\vec{r} = N_A, \int \rho_B(\vec{r})\mathrm{d}\vec{r} = N_B \quad (5)$$

where the search is performed among subsystem densities which are pure-state noninteracting $v$-representable. The sufficient condition for reaching the exact ground-state

density in SDFT is that it can be decomposed as a sum of two pure-state noninteracting $v$-representable densities comprising an integer number of electrons $N_A$ and $N_B$ (see the discussions in ref 26). FDET does not target the ground-state of the total system but the density, which minimizes the total ground-state energy in presence of the following constraint:

$$\rho \geq \rho_B \tag{6}$$

which is given in advance.

The total density obtained in FDET (eq 2 is, therefore, not equal to the exact ground-state density except for a particular case, i.e., when the difference between $\rho_o^{tot}(\vec{r})$ and the assumed $\rho_B(\vec{r})$ is representable using one of the auxiliary descriptors mentioned above: orbitals of the noninteracting reference system,[24] interacting wave function,[27] or one particle density matrix.[28] On the virtue of Hohenberg−Kohn theorems, FDET can lead only to the upper bound of the ground-state energy:

$$E_{emb}[\rho_B] \geq E_o \tag{7}$$

Any numerical implementation of FDET can be easily converted to methods solving coupled Kohn−Sham-like equations in SDFT. In fact, the first numerical implementation of SDFT applicable for molecular systems used the "freeze-and-thaw" cycle,[39] which was applied in a number of subsequent studies (see for instance refs 40−42). In the original numerical studies based on SDFT concerning atoms in solids[36,37] and in the recent numerical implementation of SDFT for molecular liquids,[43] the coupled Kohn−Sham equations are solved simultaneously. We have also shown recently that the "freeze-and-thaw" cycle can be performed simultaneously with displacing nuclear position accelerating the SDFT-based geometry optimization.[44] The "freeze-and-thaw" cycle to solve the coupled Kohn−Sham like equations is used by us in methodological studies on approximants to the bifunctional of the nonadditive kinetic potential $v_t^{nad}[\rho_A, \rho_B]$ (see for instance refs 45−47) or in preparation stages for large-scale simulations, in which the search given in eq 2 is performed for smaller model systems in order to establish the adequacy of the simplified $\rho_B(\vec{r})$ in large-scale simulations.

If a noninteracting reference system is used to perform the search given in eq 2, the corresponding orbitals ($\phi_i^A$) are obtained from the following Kohn−Sham-like equations (eqs 20−21 in ref 24):

$$\left[-\frac{1}{2}\nabla^2 + v_{eff}^{KS}[\rho_A, \vec{r}] + v_{emb}^{eff}[\rho_A, \rho_B; \vec{r}]\right]\varphi_i^A = \varepsilon_i^A \varphi_i^A \qquad i = 1, N^A \tag{8}$$

where $v_{emb}^{eff}[\rho_A, \rho_B; \vec{r}]$ is given in eq 4.

The effectiveness of methods based on eq 8 for the studying changes in the electronic structure arising due to the interactions between the embedded system and its environment was demonstrated for: vertical excitation energies,[25,48] electron spin resonance (ESR) hyperfine coupling constants,[49,50] ligand-field splittings of $f$-levels in lanthanide impurities,[51] NMR shieldings,[52] dipole and quadrupole moments, electronic excitation energies, and frequency dependent polarizabilities.[53]

In multilevel simulations based on FDET, $\rho_B$ is an assumed quantity, which can obtained following various approaches. Since the electron density is a well-defined quantity also in the macroscale, it is even possible to generate $\rho_B$ without using any quantum chemical approach.[54] A natural choice for $\rho_B$ is to use Kohn−Sham equations to generate it for the isolated environment and to use it for generating the embedding potential. Such procedure does not take into account the electronic polarization of the environment by the embedded subsystem. For this reason such density and embedding potential are labeled as nonrelaxed. If the "freeze-and-thaw" procedure[55] is used to generate $\rho_B$, which together with the optimal $\rho_A$ minimizes the energy of the whole system, then the corresponding quantities are labeled as relaxed.

Our own numerical experience concerning the applicability of eq 1 in embedding calculations shows invariably that it leads frequently to qualitatively wrong or erratic results, especially if other properties than energy are investigated (for ligand field splitting of the $f$-levels for lanthanide impurities in solids see ref 51, for redistribution of electron density in charged intermolecular complexes see ref 56, for electron density and total energy in intermolecular complexes see ref 45, for excitation energies see ref 25, for instance). The erratic results obtained using the electrostatic-only embedding potential arise from the fact that any variational method neglecting nonelectrostatic terms neither assures the proper variational limit in the sense of the second Hohenberg−Kohn theorem nor takes into account the antisymmetric character of the wave function for the whole system. The present work represents an extension of such studies. A systematic account of the flaws of the electrostatic-only embedding potential is provided for yet another quantity directly related to the electronic structure—the values of iso-g for an embedded radical molecule. The g-tensor is an important parameter of ESR spectroscopy, and its analysis is used as a supplementary tool to elucidate the protonation state of a radical in the enzyme's active center.[57] The protein provides both the steric constraints affecting the geometry of the radical as well as the electronic environment (long-range electrostatic as well as short-range Pauli repulsion), which affect the observed g-tensor. Modeling g-tensor must take a proper account of these two types of effects, and the FDET methods are especially designed to study the electronic effects (for a representative study see ref 50). The secondary aim of this study is the analysis of the errors in calculated environment-induced shifts of iso-g which can be attributed to the use of approximate density functional for the nonadditive kinetic potential $v_t^{nad}[\rho_A, \rho_B]$, i.e., the last two terms in eq 4.

We have chosen to study a model system consisting of the biliverdin IXa radical (BV) and a few amino acids of the protein phycocyanobilin (PCYA) representing the nearest neighbors of BV in the BV−PCYA complex. The g-tensor of BV has extremely small anisotropy. It should be noted that conventional ESR cannot resolve the components of g-tensors with very small anisotropy (as those of organic

radicals), but recent developments in high-field ESR allow one to determine the components of even such g-tensors and to use them in the interpretation of biochemical data.[57] It is known that PCYA catalyzes the reduction of BV to PCYA, the precursor of biliprotein chromophores found in phycobilisomes. The present work represents the preparatory stage of the project aimed at studying the whole protein−radical complex by means of multilevel type of simulations based on FDET. Two numerical model related issues: the adequacy of the used the approximant for the bifunctional of the nonelectrostatic components of the embedding potential and the adequacy of the choice for $\rho_B$ are, therefore, at the focus of the present work.

In order to obtain the absolute values of iso-g as and the environment induced shifts of this quantity [$\Delta$iso-g = iso-g(BV + environment) − iso-g(BV)], the calculations based on either embedding or supermolecular strategy are feasible for the investigated model system, which is rather small. The possible advantages of using the FDET-based embedding calculations over the supermolecular ones lie not only in reducing the computational costs. If the shifts arise from nonbonding interactions between the embedded molecule and its environment, then the quality of the shifts obtained from FDET strategy can be expected to be better than their supermolecular counterparts for the reasons addressed in more detail below. The quality of the results obtained in both the Kohn−Sham and the embedding calculations depend on such common factors as: (a) the molecular model of the real system (size of the model and coordinates of atoms not available from experiments), (b) the approximant for the exchange−correlation functional, (c) the treatment of relativistic effects (the spin−orbit coupling in particular), and (d) the basis set. As far as the environment induced shifts are concerned, however, their quality in supermolecular and embedding calculations is determined by different factors. In the supermolecular case, the shift is the difference between values obtained from two independent Kohn−Sham calculations. As a result, the errors in iso-g might cumulate or cancel. In the embedding case, the calculated shift arises from the addition of the embedding potential to the effective potential corresponding to the system without any environment. The errors in the shifts are, therefore, determined mainly by the quality of the embedding potential given in eq 4. Due to its first-principles based origin, the quality of the embedding potential used in practice is determined by only two factors: the chosen frozen electron density ($\rho_B$) and the used approximant for the nonadditive kinetic potential (the last two terms in eq 4). The importance of the first factor can be easily monitored by performing the "freeze-and-thaw" calculations on model systems. Concerning the orbital-free embedding potential expressed as the functional of $\rho_A$ and $\rho_B$, its medium- and long-range part is known exactly (electrostatics) and can be expected to dominate if the embedded molecule is surrounded by noncovalently bound polar molecules. Since the errors in the shifts of the properties related directly to the electronic structure, i.e., to the embedded orbitals and their energies, are determined only by the accuracy of the approximants to the nonadditive kinetic− and exchange−correlation potentials, one can expect that, whenever the environment includes polar or charged molecules represented exactly in the orbital-free embedding potential given in eq 4, the shifts obtained from FDET-based calculations surpass in accuracy the ones derived using their supermolecular counterparts.

The model environment of BV considered in the present work is the same as the one investigated earlier in the study aimed at determination of the protonation state of PCYA−bound BV,[57] where the supermolecular calculations were used to interpret experimental data. We use the same model for different purposes: (a) to explore the possibility of replacing supermolecular calculations by embedding ones for these types of analyses and (b) to determine the necessary conditions (choice of the frozen density, number of centers to expand the embedded orbitals using atomic bases, possibility to neglect nonelectrostatic components of the exact orbital-free embedding potential) for the optimal embedding calculations, i.e., assuring the smallest of the deviations from the supermolecular results at the largest reduction of the computational efforts.

## Computational Details

The three considered protonation states denoted with $\alpha ACD\delta^{\bullet+}$, $ACD\delta^{\bullet}$, $ACD^{\bullet-}$ are shown in Figure 1. In the label, a big letter indicates a protonated nitrogen site in the pyrrole ring (A, B, C, or D), a Greek letter ($\alpha$ or $\delta$) denotes a protonated carbonyl oxygen site, and the + or − denote the charge of the embedded species.

Two cluster models of the environment of the PCYA−bound BV shown in Figure 2 are considered: the smaller (EHD) and the larger (EHDYQ) (the amino acids are referred to with commonly used one letter codes). The EHDYQ cluster comprises amino acids: GLU76, HIS88, ASP105, TYR212, and GLN216 in their neutral form taken from the 2D1E structure deposited in the Protein Data Bank (PDB).[58] The EHD cluster comprises GLU76, HIS88, and ASP105 (all polar). These three amino acids are chosen because of their proximity to the carbonyl oxygen atoms on the pyrolle rings. All of them are capable of hydrogen bonding to the protonated BV and are thus critical for determining the locations of protons in BV. They are expected to affect the g-tensor of the protein-bound BV. The amino acids of the EHD cluster are also believed to be essential in catalysis as proton donors.[59] Concerning the larger cluster (EHDYQ), it comprises additionally to two polar amino acids TYR212 and GLN216. They are near the $\alpha$ carbonyl oxygen atom of the pyrrole ring and far from the $\delta$ carbonyl oxygen atom.

The position of the hydrogen atoms, which is not available in the PDB structure, was generated initially using "Accelrys DS Visualizer suite"[60] and subsequently refined using the nonrelativistic Kohn−Sham LDA calculations, with the STO-type triple-$\zeta$ with polarization functions basis sets (TZP label in the ADF version 2009.01 basis set library). The Cartesian coordinates are provided in the Supporting Information.

The g-tensor, the main property investigated in the present work, is related to the magnetic moment of an electron as: $\mu = -g\beta S$ where $S$ is the electron spin, and $\beta$ the bohr magneton. The $g$-value depends on the particular magnetic species under consideration.[65] The principal components of

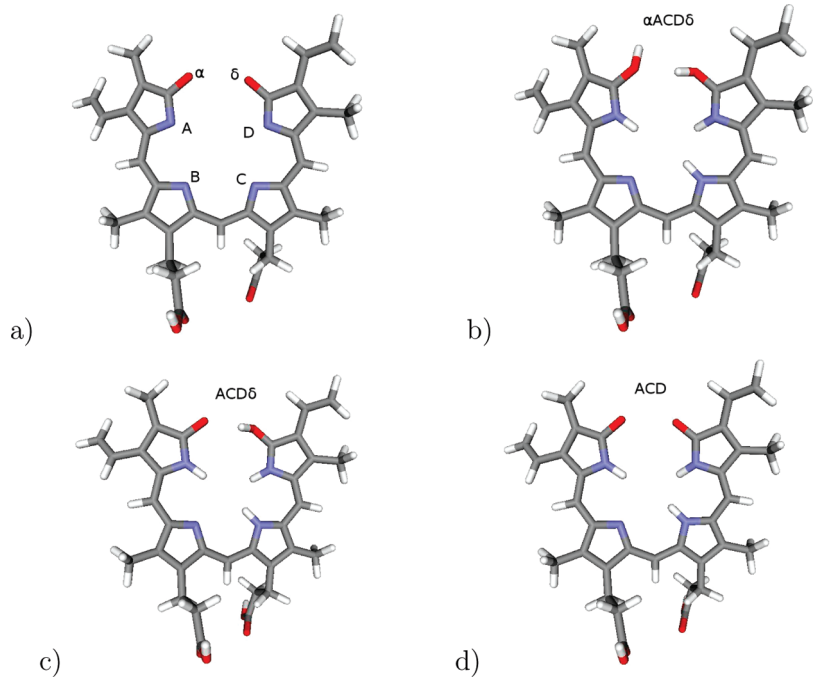



***Figure 1.*** (a) The six positions in the BV IXa radical that can be protonated: α, δ, A, B, C, and D. The three considered protonation states: (b) αACD$\delta^{\bullet+}$, (c) ACD$\delta^{\bullet}$, and d) ACD$^{\bullet-}$.

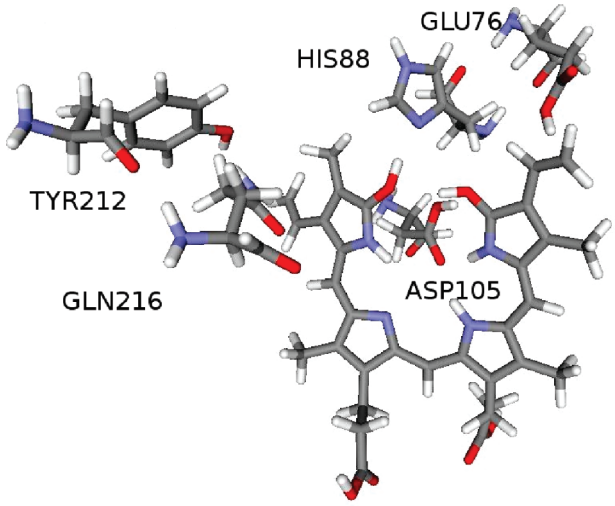



***Figure 2.*** BV IXa in the αACD$\delta^{\bullet+}$ protonation state (for the convention, see Figure 1) in ferredoxin oxidoreductase represented by five amino acids: GLU76, HIS88, ASP105, TYR212, and GLN216.

the g-tensor together with the hyperfine tensor $A$ specify the positions and the splittings of the lines in ESR spectra, depending on the direction of the magnetic field relative to the molecular axis. The spectral anisotropy is completely specified by three $g$-values ($g_{xx}$, $g_{yy}$, $g_{zz}$) and three hyperfine constants ($A_{xx}$, $A_{yy}$, $A_{zz}$). The isotropic $g$-value (iso-g), which is discussed throughout the present work, is the average of the principal $g$-values: iso-g $= 1/3(g_{xx} + g_{yy} + g_{zz})$.

In this study, the molecular $g$-values are calculated based on the solutions of the relativistic Kohn−Sham equations in the "spin−orbit coupling containing zeroth order regular approximation (ZORA) to the Dirac equation".[66−69] In ZORA, the spin−orbit coupling interaction (which is for many systems the most important factor for shifting the g-tensor components away from the free-electron value $g_e$) and other relativistic effects are taken into account variationally. The spin-restricted version of the spin−orbit including ZORA calculations,[70] which is computationally less expensive than the spin-unrestricted version[71] and thus more attractive for the study of average and big sized systems, was used. The method introduced in ref 70 combining ZORA with a single-orbital reference technique to deal with gauge dependency for open-shell doublet systems was applied.

The Becke−Perdew exchange−correlation functional[72,73] is used in both supermolecular and embedding calculations. This choice is motivated by its reported reliability in the calculation of the ESR parameters.[70]

The recently developed approximant to the nonadditive kinetic energy bifunctional (NDSD),[46] which satisfies the uniform electron gas limit and the asymptotic form of the exact bifunctional for the nonadditive kinetic potential at $\rho_A \to 0$ and $\int \rho_B d\vec{r} = 2$, was used. It takes the following form:

$$\tilde{T}_s^{\mathrm{nad(NDSD)}}[\rho_A, \rho_B] = \frac{3}{10}(3\pi^2)^{2/3} \int ((\rho_A + \rho_B)^{5/3} - \rho_A^{5/3} - \rho_B^{5/3}) d\vec{r} + \int f(\rho_B, \nabla\rho_B) \cdot \rho_A(\vec{r}) \cdot v_t^{\mathrm{limit}}[\rho_B](\vec{r}) d\vec{r} \quad (9)$$

where

$$v_t^{\mathrm{limit}}[\rho_B](\vec{r}) = \frac{1}{8}\frac{|\nabla\rho_B|^2}{\rho_B^2} - \frac{1}{4}\frac{\nabla^2\rho_B}{\rho_B} \quad (10)$$

$$f(\rho_B, \nabla\rho_B) = (\exp(\lambda(-s_B + s_B^{\min})) + 1)^{-1} \times (1 - (\exp(\lambda(-s_B + s_B^{\max})) + 1)^{-1}) \times (\exp(\lambda(-\rho_B + \rho_B^{\min})) + 1)^{-1} \quad (11)$$

and where: $s_B = (|\nabla\rho_B|)/(2(3\pi^2)^{1/3}\rho_B^{4/3})$, $s_B^{\min} = 0.3$, $s_B^{\max} = 0.9$, $\rho_B^{\min} = 0.7$, and $\lambda = 500$.

The STO-type DZP(ZORA) basis set from the ADF version 2009.01 package's basis set library,[74] that comprises a valence double-ζ basis set with one set of polarization functions, is used in the evaluation of g-tensor in both embedding and supermolecular calculations. All reported calculations are performed using the ADF version 2009.01 package.[74] Considering the atomic basis sets, it is important to point out a different significance of the choice of the number of centers used to construct molecular orbitals in supermolecular and in embedding calculations. In the supermolecular calculations, the number of centers is rarely considered an issue as they coincide with the positions of nuclei, except for some particular types of calculations. For instance, the "ghosts" are used commonly to evaluate the basis set superposition error by means of the counterpoise procedure.[61] In embedding calculations, it is natural to restrict the of centers of atomic basis sets to the nuclei of the embedded species. This is one of the sources of great computational savings in the embedding strategy after all. Addition of atomic sets localized in the environment might, however, significantly affect the numerical results of the embedding calculations.[45,46,62,63] Since eq 8 is based on variational principle (eq 2), any additional basis functions used to construct the embedded orbitals can only bring the results closer to the variational limit. The supermolecular expansion is, therefore, used as a standard in development/testing approximants for $v_t^{\mathrm{nad}}[\rho_A, \rho_B]$.[39,45,64] The use of full set of atomic centers, i.e., that in the embedded system and in the environment, is, however, not practical for large-scale simulations but is frequently applied in the preliminary stage of a simulations to chose the minimal size of the basis set and the centers assuring numerical stability for the results.[26] In the present work, the considered systems are rather small and both variants of embedding calculations are feasible. They are referred to as monomer and supermolecular expansion, respectively.

Finally, it is important to underline that to investigate the possibility of replacing the supermolecular calculations by the embedding ones using the considered embedding potentials, it is crucial that the embedding results and their supermolecular counterparts are obtained using the same: type of atomic basis sets, centers of the atomic basis sets, numerical grid, approximant for the exchange−correlation potential, and the external potential (geometry of the cluster). The particular choices for made for these quantities is of secondary importance from the point of view of the comparisons made in this work as long as the common parameters are the same in supermolecular and embedding calculations. The comparisons between the embedding and supermolecular results are made for several systems (various protonation states of BV and sizes of the environment) to ensure that the results are not accidental and to cover various possible interaction modes of a radical molecule with a protein. Nevertheless, some choices were made arbitrarily such as that for the approximant for the exchange−correlation potential, the choice of the protonation state of the amino acids, and the basis set used in most of the analysis. It is also important to underline that despite the fact that the investigated models are constructed based on the BV−PCYA complex, we do not attempt to reproduce quantitatively the experimental g-tensors because the considered models are most likely too small to be used for such purposes.

***Table 1.*** Iso-g and Δiso-g = Iso-g(BV + Environment) − Iso-g(BV) of BV in Various Environments Calculated with Eqs 1 and 4 and Nonrelaxed Density of the Environment[a]

| method[b] | environment | protonation state of BV | iso-g | Δiso-g (ppm) |
|---|---|---|---|---|
| embedding: eq 1[c] | EHD | αACDδ•+ | 2.002810 | −43 |
| embedding: eq 1[d] | EHD | αACDδ•+ | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | αACDδ•+ | 2.002905 | 52 |
| embedding: eq 4[d] | EHD | αACDδ•+ | 2.002900 | 47 |
| Kohn−Sham[c] | none | αACDδ•+ | 2.002853 | 0 |
| Kohn−Sham[d] | none | αACDδ•+ | 2.002849 | 0 |
| Kohn−Sham | EHD | αACDδ•+ | 2.002928 | 75 |
| embedding: eq 1[c] | EHDYQ | αACDδ•+ | 2.002811 | −42 |
| embedding: eq 1[d] | EHDYQ | αACDδ•+ | NC[e] | NC[e] |
| embedding: eq 4[c] | EHDYQ | αACDδ•+ | 2.002904 | 51 |
| embedding: eq 4[d] | EHDYQ | αACDδ•+ | 2.002898 | 45 |
| Kohn−Sham[d] | none | αACDδ•+ | 2.002848 | 0 |
| Kohn−Sham | EHDYQ | αACDδ•+ | 2.002982 | 129 |
| embedding: eq 1[c] | EHD | ACDδ• | 2.003639 | 50 |
| embedding: eq 1[d] | EHD | ACDδ• | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | ACDδ• | 2.003742 | 153 |
| embedding: eq 4[d] | EHD | ACDδ• | 2.003745 | 156 |
| Kohn−Sham[c] | none | ACDδ• | 2.003589 | 0 |
| Kohn−Sham[d] | none | ACDδ• | 2.003592 | 0 |
| Kohn−Sham | EHD | ACDδ• | 2.003781 | 192 |
| embedding: eq 1[c] | EHD | ACD•− | 2.004471 | −101 |
| embedding: eq 1[d] | EHD | ACD•− | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | ACD•− | 2.004667 | 95 |
| embedding: eq 4[d] | EHD | ACD•− | 2.004661 | 89 |
| Kohn−Sham[c] | none | ACD•− | 2.004572 | 0 |
| Kohn−Sham[d] | none | ACD•− | 2.004564 | 0 |
| Kohn−Sham | EHD | ACD•− | 2.004494 | −78 |

[a] Kohn−Sham results for the isolated and complexed BV are also given for reference. [b] STO-type DZP(ZORA) basis set was used. [c] Monomer basis set. [d] Supermolecular basis set. [e] Self-consistent cycle did not converge.

## Results and Discussions

Throughout the results section, the environment induced shifts of the iso-g obtained from the embedding calculations are compared with the shifts obtained from the supermolecular strategy. The absolute or relative value of the difference: $\delta\Delta\text{iso-g} = \Delta\text{iso-g}^{(\text{supermolecule})} - \Delta\text{iso-g}^{(\text{embedding})}$ is used in all discussions concerning the accuracy of various embedding potentials considered in the present work. Note that the difference δΔiso-g is equal to the difference between the absolute values of iso-g calculated in supermolecular and embedding strategies ($\delta\text{iso-}g = \text{iso-g}^{\text{supermolecule}} - \text{iso-g}^{\text{embedding}}$).

Table 1 collects the iso-g and Δiso-g values obtained using nonrelaxed $\rho_B$ for the three protonation states of BV in various environments. The reference Kohn−Sham results are also given for comparison. In three among four investigated systems, eq 1 leads to qualitatively wrong shifts (results obtained with monomer expansion). The situation is even worse when supermolecular expansion is used. The order of orbital levels is wrong, and no self-consistent solutions of eq 8 can be obtained. The situation is greatly improved if the full embedding potential given in eq 4 is used. The results of supermolecular and monomer expansion are quite similar

***Table 2.*** Iso-g and ΔIso-g = Iso-g(BV + Environment) − Iso-g(BV) of BV in Various Environments Calculated with Eqs 1 and 4 and Relaxed Density of the Environment[a]

| method[b] | environment | protonation state of BV | iso-g | Δiso-g (ppm) |
|---|---|---|---|---|
| embedding: eq 1[c] | EHD | αACDδ•+ | 2.002876 | 23 |
| embedding: eq 1[d] | EHD | αACDδ•+ | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | αACDδ•+ | 2.002919 | 66 |
| embedding: eq 4[d] | EHD | αACDδ•+ | 2.002919 | 66 |
| Kohn−Sham[c] | none | αACDδ•+ | 2.002853 | 0 |
| Kohn−Sham | EHD | αACDδ•+ | 2.002928 | 75 |
| embedding: eq 1[c] | EHDYQ | αACDδ•+ | 2.002883 | 30 |
| embedding: eq 1[d] | EHDYQ | αACDδ•+ | NC[e] | NC[e] |
| embedding: eq 4[c] | EHDYQ | αACDδ•+ | 2.002918 | 65 |
| embedding: eq 4[d] | EHDYQ | αACDδ•+ | 2.002917 | 64 |
| Kohn−Sham | EHDYQ | αACDδ•+ | 2.002982 | 129 |
| embedding: eq 1[c] | EHD | ACDδ• | 2.003629 | 40 |
| embedding: eq 1[d] | EHD | ACDδ• | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | ACDδ• | 2.003706 | 117 |
| embedding: eq 4[d] | EHD | ACDδ• | 2.003689 | 100 |
| Kohn−Sham[c] | none | ACDδ• | 2.003589 | 0 |
| Kohn−Sham | EHD | ACDδ• | 2.003781 | 192 |
| embedding: eq 1[c] | EHD | ACD•− | 2.004347 | −225 |
| embedding: eq 1[d] | EHD | ACD•− | NC[e] | NC[e] |
| embedding: eq 4[c] | EHD | ACD•− | 2.004448 | −124 |
| embedding: eq 4[d] | EHD | ACD•− | 2.004408 | −164 |
| Kohn−Sham[c] | none | ACD•− | 2.004572 | 0 |
| Kohn−Sham | EHD | ACD•− | 2.004494 | −78 |

[a] Kohn-Sham results for the isolated and complexed BV are also given for reference. [b] STO-type DZP(ZORA) basis set was used. [c] Monomer basis set. [d] Supermolecular basis set. [e] Self-consistent cycle did not converge.

in all cases. The complexation induced shifts obtained from embedding calculations with eq 4 reproduce qualitatively the reference values from supermolecular calculations in three considered cases but lead to qualitatively wrong shifts for $ACD^{\bullet-}$. According to the discussion in the Introduction, these discrepancies might originate from the choice of $\rho_B$ or from the fact that an approximated functional was used for $v_t^{nad}[\rho_A, \rho_B]$. We note, however, that for embedded species which are electrically neutral, the agreement between the reference data and the shifts from the embedding calculations using the potential given in eq 4 is the best (156 vs 192 ppm). This indicates that the lack of relaxation might lie at the origin of the large discrepancies in the other cases.

To investigate this interpretation, the shifts were evaluated using the relaxed $\rho_B$ which is obtained in the "freeze-and-thaw" procedure. They are collected in Table 2 which parallels Table 1. The relaxed $\rho_B$ takes into account the electronic polarization of the environment by the embedded species and corresponds to the variational minimum of the total energy functional expressed using the applied approximated density functionals (SDFT). On the practical side, we notice that the "freeze-and-thaw" procedure converges rapidly and that the shifts do not change noticeably after completion of just one cycle (data not shown). Let us start with the worst case analyzed previously, i.e., $ACD^{\bullet-}$. Indeed, the qualitatively wrong complexation induced shift obtained in the absence of relaxation (+89 vs −78 ppm reference value) becomes negative (−164 ppm). The embedding potential of eq 4 leads to the correct direction of the effect but overestimates its magnitude. Removal of the basis functions localized in the environment brings the numerical result even closer to the supermolecular reference (−124 ppm). Similarly as in the nonrelaxed $\rho_B$ case, the differences between the shifts obtained with monomer and supermolecular basis sets are significantly smaller than the magnitudes of the shifts. The situation is quite different if the electrostatic-only (eq 1) embedding potential is used. Similarly as in the previously considered nonrelaxed case, it can be noticed that the electrostatic-only embedding potential cannot be applied with supermolecular basis sets. The results obtained using the monomer expansion with the electrostatic-only embedding potential (the situation most closely resembling the typical QM/MM computational protocol) are reasonable, but they are systematically worse than the ones obtained using the full embedding potential.

***Table 3.*** Basis Set Dependence of Iso-g and ΔIso-g = Iso-g(BV + EHD) − Iso-g(BV) Obtained in Embedding Calculations and Supermolecular (Kohn−Sham) Calculations

| method[a] | basis set | iso-g | Δiso-g (ppm) |
|---|---|---|---|
| Kohn−Sham[c] | STO(DZ) | 2.002959 | 72 |
| Kohn−Sham[c] | STO(DZP) | 2.002928 | 75 |
| embedding: eq 1[b,c] | STO(DZ) | 2.002871 | −16 |
| embedding: eq 1[b,c] | STO(DZ+ghosts) | NC[d] | NC[d] |
| embedding: eq 1[b,c] | STO(DZP) | 2.002810 | −43 |
| embedding: eq 1[b,c] | STO(DZP+ghosts) | NC[d] | NC[d] |
| embedding: eq 1[b,d] | STO(DZ) | 2.002922 | 35 |
| embedding: eq 1[b,d] | STO(DZ+ghosts) | NC[d] | NC[d] |
| embedding: eq 1[b,d] | STO(DZP) | 2.002876 | 23 |
| embedding: eq 1[b,d] | STO(DZP+ghosts) | NC[d] | NC[d] |
| embedding: eq 4[b,c] | STO(DZ) | 2.002946 | 59 |
| embedding: eq 4[b,c] | STO(DZ+ghosts) | 2.002938 | 51 |
| embedding: eq 4[b,c] | STO(DZP) | 2.002905 | 52 |
| embedding: eq 4[b,c] | STO(DZP+ghosts) | 2.002900 | 47 |
| embedding: eq 4[b,d] | STO(DZ) | 2.002959 | 72 |
| embedding: eq 4[b,d] | STO(DZ+ghosts) | 2.002954 | 67 |
| embedding: eq 4[b,d] | STO(DZP) | 2.002919 | 66 |
| embedding: eq 4[b,d] | STO(DZP+ghosts) | 2.002919 | 66 |

[a] STO-type DZP(ZORA) basis set was used. [b] Results for BV in the αACDδ•+ protonation state. [c] Nonrelaxed $\rho_B$. [d] Relaxed $\rho_B$.

As far as the basis set effect on the obtained complexation induced shifts is concerned, it is worthwhile to notice that the two previously considered choices for the basis sets (monomer or supermolecular expansion) represent two extremes in practical calculations. Most of the QM/MM calculations do not employ the basis sets localized in the environment. From the practical point of view, however, it is more useful to analyze the numerical stability of the shifts obtained from embedding calculations if only the monomer expansion is considered but the atomic basis sets change. The following section addresses this issue for one case: BV in the $\alpha ACD\delta^{\bullet+}$ protonation state in the EHD environment.

We start with noticing that, the complexation induced shifts of iso-g obtained from supermolecular calculations are not affected by the change of the basis set significantly (see Table 3). The basis set superposition error on Δiso-g obtained from the supermolecular strategy is negligible reaching 3 ppm at most (data not shown). Embedding calculations using Coulombic embedding potential appear to be very sensitive to the choice of the basis set. With the monomer expansion, the shifts equal to 35 and 23 ppm for STO-DZ and STO-DZP, respectively. These values represent less than 50% of the reference shifts (72 or 75 ppm). Reducing the size of the basis set on each atom in the supermolecular expansion

form STO-DZP to STO-DZ does not bring solution to the problem of the lack of convergence in the electrostatic-only embedding case.

The shifts obtained from embedding calculations with the full embedding potential are significantly more stable numerically and closer to the reference. With the monomer expansion the corresponding shifts are 72 and 66 ppm, which are also very close to the corresponding supermolecular reference (72 and 75 ppm for STO−DZ and STO−DZP, respectively). Moreover, adding the centers localized in the environment does not affect the obtained shifts significantly. The shifts obtained with four considered sets of basis functions are scattered between 66 and 72 ppm.

The numerical results discussed so far show invariably that the full embedding potential is indispensable for obtaining numerically stable values of complexation induced shifts. However, an important factor in embedding calculations applying the potential given in eq 4 is the choice of the electron density $\rho_B$. Performing the full "freeze-and-thaw" procedure (SDFT) is not practical in the case of large environments, and a nature of multilevel simulations is that additional approximations are introduced for $\rho_B$, as in methods based on FDET. In the following part, we focus on the ACD$^{\bullet-}$ case, where nonrelaxed electron density was shown to be a very inappropriate choice for $\rho_B$, as it leads to a wrong sign of the complexation induced shift of iso-g. Moreover, the difference between the results obtained using the monomolecular and supermolecular expansion in embedding calculations was the largest as well as the effect of relaxation was also the largest in magnitude. We notice that opposite to other complexes considered in the present work, the embedded system was charged, and it was charged negatively. This suggests attribution of the failure of nonrelaxed embedding calculations to the neglect of electronic polarization and to the role of charge transfer between the embedded radical and the environment or even some covalent character of the bonding between the embedded species and its environment. The analysis of the complexation induced dipole moments in this systems supports the attribution of the failure of nonrelaxed $\rho_B$ embedding calculations to the missing electronic polarization (see Table 3). To support this interpretation of the inadequacy of choosing nonrelaxed $\rho_B$ for negatively charged environment of BV, the complexation induced dipole moments collected in Table 4 are analyzed below. The complexation induced shifts are the smallest (2.93 D) if both the embedded species and the environment are not charged, i.e., for ACD$\delta^{\bullet}$. In this case, the embedding calculations reproduce quite reasonably the complexation induced increase of the dipole moment (2.46 vs 2.93 D) but only if relaxation is taken into account. It is worthwhile to notice that despite such under performance in reproducing the complexation induced dipole moments in the absence of relaxation, the shifts of iso-g were very accurately reproduced (153 vs 192 ppm see Table 1) even if nonrelaxed $\rho_B$ is used in embedding calculations. This indicates that the quality of the embedding potential does not affect all properties of the embedded species in the same way. The choice of the simplified method to obtain $\rho_B$ in large scale simulations using eq 4 must, therefore, be subject of dedicated studies depending on the system and the investigated property. Interestingly, the complexation induced dipole moments are only slightly larger for the two cases of cationic embedded species is (3.07 and 3.94 D). Turning back to the ACD$\delta^{\bullet-}$ case, i.e., where the choice of nonrelaxed $\rho_B$ was qualitatively wrong for evaluation of $\Delta$iso-g, we notice that the complexation induced dipole moment is, indeed, the largest (7.10 D), and effect of the relaxation of $\rho_B$ on the calculated dipole moment is the largest.

***Table 4.*** Complexation Induced Dipole Moments ($|\Delta\vec{\mu}| = |\vec{\mu}_{BV+env.} - \vec{\mu}_{BV} - \vec{\mu}_{env.}|$) from Embedding and Supermolecular Calculations

| method[a] | environment | protonation state of BV | $\lvert\Delta\vec{\mu}\rvert$ (D) |
|---|---|---|---|
| Kohn−Sham | EHD | αACD$\delta^{\bullet+}$ | 3.07 |
| embedding: eq 4[b] | EHD | αACD$\delta^{\bullet+}$ | 0.73 |
| embedding: eq 4[c] | EHD | αACD$\delta^{\bullet+}$ | 2.18 |
| Kohn−Sham | EHDYD[b] | αACD$\delta^{\bullet+}$ | 3.94 |
| embedding: eq 4[b] | EHDYD | αACD$\delta^{\bullet+}$ | 1.16 |
| embedding: eq 4[c] | EHDYD | αACD$\delta^{\bullet+}$ | 1.77 |
| Kohn−Sham | EHD | ACD$\delta^{\bullet}$ | 2.93 |
| embedding: eq 4[b] | EHD | ACD$\delta^{\bullet}$ | 0.93 |
| embedding: eq 4[c] | EHD | ACD$\delta^{\bullet}$ | 2.46 |
| Kohn−Sham | EHD | ACD$^{\bullet-}$ | 7.10 |
| embedding: eq 4[b] | EHD | ACD$^{\bullet-}$ | 1.27 |
| embedding: eq 4[c] | EHD | ACD$^{\bullet-}$ | 5.87 |

[a] STO-type DZP(ZORA) basis set was used. [b] Nonrelaxed $\rho_B$. [c] Relaxed $\rho_B$.

## Conclusions

The applicability of the embedding strategy as an alternative to supermolecular calculations was investigated using a small model system comprising BV and a few amino acids of the PCYA protein. The numerical results indicate clearly that the nonelectrostatic components of the embedding potential are indispensable. Without a nonelectrostatic component of the exact embedding potential, the results of embedding calculations are not reliable as the effect of the environment on the investigated property (shift of iso-g) is described erratically (strong dependence of the results on the basis sets) and even qualitatively wrong. The magnitude of the change in iso-g resulting from changing the protonation state of the radical can be significant as estimated in this work and in the work by Stoll et al.[57] (a few hundreds of ppm). Such effects on iso-g are significantly larger in magnitude than the differences between supermolecular and embedding results. Even larger effects on the magnitude of iso-g can be expected to rise from the effect of protein on the conformation of the radical. Both supermolecular or embedding strategies can be, therefore, used to target these larger effects. The present work indicates that if the nonelectrostatic terms are taken into account by approximate density functionals, then the embedding strategy can be considered as a numerically efficient alternative. It is important to underline that neither supermolecular nor embedding calculations are likely to be used for interpretations of very small effects, such as the shift of iso-g of BV resulting from protein mutations which do not exceed 15 ppm.[57] The effect is too small to be reliably predicted by any type of calculations at least for the mutants for which the data is available.

Concerning practical calculations based on the frozen-density embedding theory,[24−28] the analytic form of the nonelectrostatic components of the exact embedding potential is unfortunately not known. Approximants must be used instead. The present work shows that the currently known approximants to these terms perform quite reasonably. The embedding calculations can reproduce the 50−90% of the reference shifts obtained using the supermolecular strategy. The fact that the shifts are not perfectly reproduced indicates that further improvements of the approximants for the kinetic energy-dependent component of the embedding potential are needed. The present work provides also important hints for setting up a large-scale multilevel simulation based on FDET. For the neutral and cationic amino acids interacting with BV, using the nonpolarized electron density to evaluate the orbital-free embedding potential is a reasonable approximation. For negatively charged amino acids, however, the electronic polarization of the environment by the embedded molecule must be taken into account. We underline that the embedding calculations using the orbital-free embedding potential lead to results which are numerically robust and do not vary significantly with the basis sets, even if additional basis sets centered on atoms in the environment are used in construction of embedded orbitals.

Finally, we stress that the practical recommendations and conclusions emerging from the present studies concerning: weak basis set dependence, negligible effect of atomic basis sets localized in the environment, reasonably good approximation of neglecting electronic polarization of neutral and cationic amino acids in the environment, adequacy of the NDSD approximant for the nonadditive kinetic energy were drawn based on the analyses of shifts of the iso-g. Investigation of other properties, the quality of which is directly linked to the accuracy of the used embedding potential, might lead to different recommendations. For instance, the relative errors (deviation from the reference supermolecular results) in the complexation induced dipole moments obtained from embedding calculations do not correlate well with the errors in the iso-g shifts despite the fact that both errors are determined by the quality of the used orbital-free embedding potential. According to our numerical experience, the complexation induced dipole moments are very sensitive to the variations in the embedding potential. This is one of the reasons we use induced dipoles in the methodological studies aimed at development or validation of approximants for the nonadditive kinetic potential.[45−47]

**Acknowledgment.** This work was supported by the grant 200020-124817 from the Fonds National Suisse de la Recherche Scientifique.

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

## References

(1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(2) Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523.

(3) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.

(4) Sauer, J.; Ugliengo, P.; Garrone, E.; Saunders, V. R. *Chem. Rev.* **1994**, *94*, 2095.

(5) Chalmet, S.; Rinaldi, D.; Ruiz-Lopez, M. F. *Int. J. Quantum Chem.* **2001**, *84*, 559.

(6) Sokol, A.; Bromley, S.; French, S.; Catlow, C.; Sherwood, P. *Int. J. Quantum Chem.* **2004**, *99*, 695.

(7) Gao, J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1995; Vol. 7; pp 119−185.

(8) Ellis, D.; Warschkow, O. *Coord. Chem. Rev.* **2003**, *238*, 31.

(9) Savin, A.; Wesolowski, T. A. *Progress in Theoretical Chemistry and Physics* **2009**, *19*, 327.

(10) Neaton, J. B.; Ashcroft, N. W. *Nature* **1999**, *400*, 141.

(11) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715.

(12) Kastner, J.; Thiel, S.; Senn, H. M.; Sherwood, P.; Thiel, W. *J. Chem. Theory Comput.* **2007**, *3*, 1064.

(13) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442.

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

(15) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem.* **1998**, *102*, 4714.

(16) Thiel, W. In *Multiscale Simulation Methods in Molecular Sciences*; Grotendorst, J., Attig, N., Blugel, S., Marx, D., Eds.; Institute for Advanced Simulation, Forschungszentrum Julich: Julich, Germany, 2009; NIC Series; Vol. 42; pp 203−214.

(17) Titmuss, S. J.; Cummins, P. L.; Rendell, A. P.; Bliznyuk, A. A.; Gready, J. E. *J. Comput. Chem.* **2002**, *23*, 1314.

(18) Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R. *J. Chem. Phys.* **2002**, *117*, 10535.

(19) V. Guallar, M.-H. B.; Lippard, S. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6998.

(20) Biswas, P. K.; Gogonea, V. *J. Chem. Phys.* **2005**, *123*, 164114.

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

(22) Tu, Y.; Laaksonen, A. *J. Chem. Phys.* **1999**, *111*, 7520.

(23) Tunon, I.; Martins-Costa, M. T.; Millot, C.; Ruiz-Lopez, M. F. *Chem. Phys. Lett.* **1995**, *241*, 450.

(24) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050.

(25) Wesolowski, T. A. *J. Am. Chem. Soc.* **2004**, *126*, 11444.

(26) Wesolowski, T. A. In *Computational Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific: Singapore, 2006; Vol. X; pp 1−82.

(27) Wesolowski, T. A. *Phys. Rev. A* **2008**, *77*, 012504.

(28) Pernal, K.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2009**, *109*, 2520.

(29) Stefanovich, E. V.; Truong, T. N. *J. Chem. Phys.* **1996**, *104*, 2946.

(30) Govind, N.; Wang, Y. A.; Carter, E. A. *J. Chem. Phys.* **1999**, *110*, 7677.

(31) Neugebauer, J.; Baerends, E. J. *Phys. Rev. A: At., Mol., Opt. Phys.* **2006**, *110*, 8786.

(32) Hodak, M.; Lu, W.; Bernholc, J. *J. Chem. Phys.* **2008**, *128*, 014101.

(33) Gomes, A. S. P.; Jacob, C. R.; Visscher, L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5353.

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

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

(36) Cortona, P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *44*, 8454.

(37) Senatore, G.; Subbaswamy, K. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *34*, 5754.

(38) Elliott, P.; Cohen, M. H.; Wasserman, A.; Burke, K. *J. Chem. Theory and Comput.* **2009**, *5*, 827.

(39) Wesolowski, T. A.; Chermette, H.; Weber, J. *J. Chem. Phys.* **1996**, *105*, 9182.

(40) Wesolowski, T. A.; Tran, F. *J. Chem. Phys.* **2003**, *118*, 2072.

(41) Kevorkyants, R.; Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2006**, *124*, 024104.

(42) Dulak, M.; Kaminski, J. W.; Wesolowski, T. A. *J. Chem. Theory Comput.* **2007**, *3*, 735.

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

(44) Dulak, M.; Kaminski, J. W.; Wesolowski, T. A. *Int. J. Quant. Chem.* **2009**, *109*, 1886.

(45) Wesolowski, T. A.; Weber, J. *Int. J. Quantum Chem.* **1997**, *61*, 303.

(46) Lastra, J. M. G.; Kaminski, J. W.; Wesolowski, T. A. *J. Chem. Phys.* **2008**, *129*, 074107.

(47) Bernard, Y. A.; Dulak, M.; Kaminski, J. W.; Wesolowski, T. A. *J. Phys. A* **2008**, *41*, 0553902.

(48) Fradelos, G.; Kaminski, J. W.; Wesolowski, T. A.; Leutwyler, S. *J. Phys. Chem. A* **2009**, *19*, 9766.

(49) Wesolowski, T. A. *Chem. Phys. Lett.* **1999**, *311*, 87.

(50) Neugebauer, J.; Louwerse, M. J.; Belanzoni, P.; Wesolowski, T. A.; Baerends, E. J. *J. Chem. Phys.* **2005**, *123*, 114101.

(51) Zbiri, M.; Atanasov, M.; Daul, C.; Garcia-Lastra, J. M.; Wesolowski, T. A. *Chem. Phys. Lett.* **2004**, *397*, 441.

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

(53) Jacob, C. J.; Neugebauer, J.; Jensen, L.; Visscher, L. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2349.

(54) Kaminski, J. W.; Gusarov, S.; Kovalenko, A.; Wesolowski, T. A. *J. Phys. Chem. A* **2010**, *114*, 6082.

(55) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71.

(56) Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2006**, *124*, 164101.

(57) Stoll, S.; Gunn, A.; Brynda, M.; Sughrue, W.; Kohler, A. C.; Ozarowski, A.; Fisher, A. J.; Lagarias, J. C.; Britt, R. D. *J. Am. Chem. Soc.* **2009**, *131*, 1986.

(58) Hagiwara, Y.; Sugishima, M.; Takahashi, Y.; Fukuyama, K. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 27.

(59) Tu, S. L.; Gunn, A.; Toney, M. D.; Britt, R. D.; Lagarias, J. C. *J. Am. Chem. Soc.* **2004**, *126*, 8682.

(60) *Accelrys DS Visualizer*, v2.0.1.7347; Accelrys Software Inc.: San Diego, CA, 2007.

(61) Boys, S. F.; Bernardi, F. *Mol. Phys.* **2002**, *100*, 65.

(62) Dulak, M.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2005**, *101*, 543.

(63) Jacob, C. R.; Wesolowski, T. A.; Visscher, L. *J. Chem. Phys.* **2005**, *123*, 174104.

(64) Wesolowski, T. A. *J. Chem. Phys.* **1997**, *106*, 8516.

(65) Knowles, P., Marsh, D., Rattle, H. *Magnetic Resonance of Biomolecules*; Wiley:London, U.K., 1976.

(66) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.

(67) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783.

(68) van Lenthe, E.; van Leeuwen, R.; Baerends, E.; Snijders, J. *Intl. J. Quantum Chem.* **1996**, *57*, 281.

(69) van Lenthe, E.; Snijders, J.; Baerends, E. *J. Chem. Phys.* **1994**, *105*, 6505.

(70) van Lenthe, E.; van der Avoird, A.; Wormer, P. *J. Chem. Phys.* **1997**, *107*, 2488.

(71) van Lenthe, E.; van der Avoird, A.; Wormer, P. *J. Chem. Phys.* **1998**, *108*, 4783.

(72) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(73) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(74) *ADF 2009 suite of programs*; Theoretical Chemistry Department, Vrije Universiteit: Amsterdam, The Netherlands; http://www.scm.com. Accessed October 21, 2010.

CT100415H

JCTC Journal of Chemical Theory and Computation

# Harmonic Vibrational Analysis in Delocalized Internal Coordinates

Frank Jensen* and David S. Palmer†

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

Received August 18, 2010

**Abstract:** It is shown that a principal component analysis of a large set of internal coordinates can be used to define a nonredundant set of delocalized internal coordinates suitable for the calculation of harmonic vibrational normal modes. The selection of internal coordinates and the principal component analysis provide large degrees of freedom in extracting a nonredundant set of coordinates, and thus influence how the vibrational normal modes are described. It is shown that long-range coordinates may be especially suitable for describing low-frequency global deformation modes in proteins.

## Introduction

The nuclear motion of atoms in a molecule can, within the Born−Oppenheimer approximation, be determined by solving the nuclear Schrödinger equation, where the solution to the electronic Schrödinger equation plays the role of a potential energy surface (PES). Near a minimum on the PES, the potential energy can be approximated by a second-order Taylor expansion, and solutions to the nuclear Schrödinger equation can be expressed as a product of solutions to the one-dimensional harmonic oscillator problem. If the quantum aspects of the nuclei are ignored, the motion of the nuclei can be determined by solving the corresponding classical vibrational equation. Harmonic vibrational analyses are, in many cases, of sufficient accuracy for identifying unknown species by comparing calculated vibrational spectra with experimentally observed ones and also form the starting point for more advanced treatment of anharmonic vibrations.[1]

The eigenvalues from the solution of the harmonic vibrational equation are proportional to the vibrational frequencies, and the associated eigenvectors are the normal coordinates, often called normal modes. The vibrational frequencies contain information regarding the curvature of the underlying PES, and it is commonly assumed that the directions of the normal modes associated with the lowest vibrational frequencies can be used to search for low-energy transition states connecting to other low-energy minima on the PES.[2] When applied to biomolecules, it has been observed that low-frequency normal modes in many cases can be used to rationalize even large-scale domain movements in proteins.[3−9] More recently, such low-frequency normal modes have been used to bias molecular dynamics simulations, with the purpose of simulating conformational transitions that otherwise would be too slow to be computationally feasible.[10] We note in passing that many of these conclusions are based on normal modes calculated by elastic network models,[9] rather than atomistic force fields, but there is substantial evidence that the two types of normal modes are in reasonably good agreement with each other.[11]

While the (tangential) direction of a normal mode at a stationary point is independent of the coordinates used for expressing the vibrational problem, the atomic coordinates for any finite displacement along a normal mode will depend on the coordinates used for solving the vibrational equation.[12,13] Cartesian coordinates have been used almost exclusively, since the vibrational problem in this case can be formulated as a simple diagonalization of a mass-weighted force constant matrix. When applied to systems with hundreds or thousands of atoms, low-frequency modes often involve a large fraction of all of the coordinates, and the degree of localization/delocalization of a vibrational mode can be quantified in terms of a localization or collectivity index.[14] The vibrational problem, however, can be solved in any nonredundant set of coordinates,[15] and in the present paper, we explore whether it is possible to select other sets of coordinates that allow a more concise description of, in particular, the low-frequency vibrations in biomolecules like proteins. For small

* Corresponding author. E-mail: frj@chem.au.dk.

† Current address: Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, DE-04103 Leipzig, Germany.

10.1021/ct100463a 
Published on Web 12/03/2010

systems, there is a substantial amount of work reported for quantum vibrational analyses in internal coordinates,[16] while the applications to larger systems are significantly more scarce. Kamiya et al. have discussed how to efficiently calculate the second derivative matrix in internal coordinates from a parametrized energy function and use this to perform vibrational analysis in selected sets of internal coordinates.[17] Transformation from Cartesian coordinates to internal coordinates for analysis purposes has also been discussed.[18,19]

## Theory

The determination of vibrational normal coordinates from a set of general coordinates goes back to the seminal work of Wilson et al.,[20,21] where a set of $3N - 6$ nonredundant internal coordinates were assumed to be available. Fogarasi et al. introduced the use of natural internal coordinates for geometry optimization,[22] and several other groups have generalized this idea to extract $3N - 6$ nonredundant linear combinations, often denoted delocalized internal coordinates,[23] from a large set of redundant primitive internal coordinates.[24−28] We will in the present paper use these concepts to determine vibrational normal coordinates as a linear combination of delocalized internal coordinates extracted from a (large) set of primitive internal coordinates. Lower case bold symbols like **a** will, in the following, denote a vector with a length given by the number of variables, while a capital symbol like **A** denotes a matrix containing all **a** vectors as columns.

We will assume a nonlinear system with $N$ atoms and $3N$ Cartesian coordinates contained in a vector **x**, and thus $3N - 6$ vibrational degrees of freedom. In Cartesian coordinates, the classical kinetic and potential energy operators can be written as in eqs 1−3,[29] where the harmonic approximation has been employed for the potential energy and dotted vectors indicate time derivatives.

$$2\mathbf{T} = \sum_i^N \dot{\mathbf{x}}_i m_i \dot{\mathbf{x}}_i = \dot{\mathbf{x}}^t \mathbf{M} \dot{\mathbf{x}} \tag{1}$$

$$2\mathbf{V} = \sum_{ij}^N \Delta\mathbf{x}_i F_{ij} \Delta\mathbf{x}_j = \Delta\mathbf{x}^t \mathbf{F} \Delta\mathbf{x} \tag{2}$$

$$F_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j};\ M_{ij} = \delta_{ij} M_i \tag{3}$$

The system can alternatively be described by a set of $K$ internal coordinates **q**, where $K \geq 3N - 6$. The internal coordinates will typically be bond stretch coordinates (distance between two atoms), bending coordinates (angle between three atoms), and torsional coordinates (dihedral angle defined by four atoms) but may also be other types of coordinates.[30] A set of all possible stretch (S), bend (B), and torsion (T) coordinates can be generated automatically from a set of Cartesian coordinates by employing suitable atomic radii for determining the bonding pattern,[27,28] and such a set will in the following be denoted the primitive coordinates.

Each internal coordinate $q_i$ is a function of the Cartesian coordinates and can be Taylor expanded as shown in eq 4.

$$q_i = q_i(\mathbf{x}) = q_i(\mathbf{x}_0) + \frac{\partial q_i}{\partial x}(\mathbf{x} - \mathbf{x}_0) + \ldots \tag{4}$$

At a stationary point, only the first derivative of the internal coordinates with respect to Cartesian coordinates is required for transforming the kinetic and potential energy operators to internal coordinates,[12,31,32] which can be written in terms of the Wilson **B** matrix and the corresponding (generalized) inverse $\mathbf{B}^{-1}$ matrix.

$$\Delta\mathbf{q} = \mathbf{B}\Delta\mathbf{x} \tag{5}$$

$$B_{ij} = \frac{\partial q_i}{\partial x_j} \tag{6}$$

$$\Delta\mathbf{x} = \mathbf{B}^{-1}\Delta\mathbf{q} \tag{7}$$

$$\mathbf{B}^{-1} = \mathbf{B}^t(\mathbf{B}\mathbf{B}^t)^{-1} = \mathbf{B}^t\mathbf{G}^{-1} \tag{8}$$

The vibrational part of the **B** matrix is rectangular with dimension $(3N - 6) \times 3N$, where the remaining six vectors describing overall translation and rotation can be determined by the Eckart conditions.[33] The kinetic and potential energies transformed to internal coordinates are shown in eqs 9−11.

$$2\mathbf{T} = \dot{\mathbf{q}}^t((\mathbf{B}^{-1})^t\mathbf{M}\mathbf{B}^{-1})\dot{\mathbf{q}} = \dot{\mathbf{q}}^t\mathbf{G}_\mathbf{m}^{-1}\dot{\mathbf{q}} \tag{9}$$

$$\mathbf{G}_\mathbf{m} = \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^t \tag{10}$$

$$2\mathbf{V} = \Delta\mathbf{q}^t((\mathbf{B}^{-1})^t\mathbf{F}\mathbf{B}^{-1})\Delta\mathbf{q} = \Delta\mathbf{q}^t\mathbf{F}_\mathbf{q}\Delta\mathbf{q} \tag{11}$$

Determining the vibrational normal coordinates corresponds to choosing a linear transformation **L** that simultaneously makes $\mathbf{G_m}^{-1}$ a unit matrix and $\mathbf{F_q}$ a diagonal matrix, which is equivalent to finding the eigenvectors of the $\mathbf{G_m F_q}$ matrix product.[20,21]

$$\mathbf{G_m F_q L} = \mathbf{L}\boldsymbol{\Lambda} \tag{12}$$

Since the $\mathbf{G_m}$ matrix is nondiagonal, the $\mathbf{G_m F_q}$ matrix is nonsymmetric, but the problem can be symmetrized by transforming the **l** coordinates by $\mathbf{G_m}^{-1/2}$.[34]

$$(\mathbf{G_m}^{1/2}\mathbf{F_q}\mathbf{G_m}^{1/2})(\mathbf{G_m}^{-1/2}\mathbf{L}) = (\mathbf{G_m}^{-1/2}\mathbf{L})\boldsymbol{\Lambda} \tag{13}$$

$$(\boldsymbol{G}_\mathrm{m}^{1/2}\boldsymbol{F}_\mathrm{q}\boldsymbol{G}_\mathrm{m}^{1/2})\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \tag{14}$$

$$\boldsymbol{L} = \boldsymbol{G}_\mathrm{m}^{1/2}\boldsymbol{U} \tag{15}$$

The vibrational normal coordinates are defined by the eigenvectors obtained by diagonalization of the $\mathbf{G_m}^{1/2}\mathbf{F_q}\mathbf{G_m}^{1/2}$ matrix, where the square root of the eigenvalues $\lambda$ is proportional to the vibrational frequencies.[35] The normal coordinates defined by the **u** vectors can be back-transformed to **l** coordinates by multiplying by $\mathbf{G_m}^{1/2}$, where they are given as a linear combination of internal coordinates **q**.

$$\mathbf{l_k} = \sum_i^\mathrm{K} c_{ki} q_i \tag{16}$$

The square of the coefficients $c_{ki}$ determines the contribution of the internal coordinate $q_i$ to the (normalized) $\mathbf{l_k}$ normal

mode. The normal modes can be further transformed to Cartesian coordinates by eq 7, which must be done iteratively as the $\mathbf{B}^{-1}$ matrix only describes the first-order change.[36] We note that the normal modes are orthogonal in the **u** space, but not in the **l** or **x** spaces.

For small systems ($N < \sim 100$ atoms), it is usually possible to select a nonredundant set ($K = 3N - 6$) of primitive coordinates manually. The requirement that the **q** coordinates are nonredundant is equivalent to requiring that the $\mathbf{G_m}$ matrix is nonsingular. In practical applications, a near-singular $\mathbf{G_m}$ must also be avoided for numerical reasons, and a heuristic criterion in terms of its determinant shown in eq 17 has been proposed.[28]

$$-\log(|\mathbf{G_m}|) < \sim(3N - 6) \tag{17}$$

It should be noted that this criterion is rather conservative, as satisfactory numerical accuracy can be obtained for substantially smaller values of the $\mathbf{G_m}$ determinant; for example, **q** coordinates corresponding to $-\log(|\mathbf{G_m}|) \sim 3(3N - 6)$ can be handled without a significant loss of numerical accuracy. The determinant of $\mathbf{G_m}$ is closely related to the condition number, which has been used by Németh et al. as a criterion for selecting primitive internal coordinates in connection with geometry optimization.[37]

For systems with more than ~100 atoms, it is often difficult to define a set of primitive coordinates that does not lead to a near singular $\mathbf{G_m}$ matrix. A redundant set of internal coordinates ($K > 3N - 6$), however, may be transformed into $3N - 6$ linearly independent and $K - (3N - 6)$ dependent coordinates by a transformation matrix **S** (assuming that the primitive coordinates span the full $3N - 6$ vibrational space). The **S** matrix can be chosen to reflect the molecular symmetry (if any) and is therefore often called a symmetry matrix. The **S** matrix has dimension $K \times K$, but only the $3N - 6$ linearly independent coordinates are of interest; i.e., only the $(3N - 6) \times K$ elements of the **S** matrix are significant. We will denote the symmetry coordinates by **d**, and the transformation can be written as in eq 18.

$$\mathbf{d} = \mathbf{S}^t\mathbf{q} \tag{18}$$

A change in the nonredundant symmetry coordinates can be transformed to a change in the primitive coordinates, as shown in eq 19.

$$\Delta\mathbf{q} = \mathbf{S}\Delta\mathbf{d} \tag{19}$$

The vibrational analysis in eqs 9−15 carries directly over to symmetry coordinates, where the $3N - 6$ symmetry coordinates **d** replace the $3N - 6$ internal coordinates **q**. As the symmetry coordinates are just fixed linear combinations of the internal coordinates, the change corresponds to replacing the **B** and $\mathbf{B}^{-1}$ matrices by their symmetry transformed equivalent $\mathbf{B}'$ and $\mathbf{B}'^{-1}$ matrices.

$$\Delta\mathbf{d} = \mathbf{S}^t\Delta\mathbf{q} = \mathbf{S}^t\mathbf{B}\Delta\mathbf{x} = \mathbf{B}'\Delta\mathbf{x} \tag{20}$$

$$\Delta\mathbf{x} = \mathbf{B}^{-1}\Delta\mathbf{q} = \mathbf{B}^{-1}\mathbf{S}\Delta\mathbf{d} = \mathbf{B}'^{-1}\Delta\mathbf{d} \tag{21}$$

The corresponding $\mathbf{G_m}$ and $\mathbf{F_q}$ matrices expressed in symmetry coordinates will be denoted $\mathbf{G'_m}$ and $\mathbf{F_d}$.

There is considerable freedom in choosing the symmetry matrix, but it should be chosen such that the resulting $\mathbf{G'_m}$ matrix is sufficiently nonsingular. One convenient way of ensuring this is to define the symmetry matrix by the eigenvectors of the **G** matrix in eq 22.

$$\mathbf{G} = \mathbf{B}\mathbf{B}^t \tag{22}$$

The **G** matrix in eq 22 is equivalent to the $\mathbf{G_m}$ matrix in eq 10 without the mass weighting and also enters the expression for the $\mathbf{B}^{-1}$ matrix, eq 8. If the **b** vectors used for constructing **G** in eq 22 are taken to be normalized, this is equivalent to performing a principal component analysis (PCA) of the **b** vectors.[15,38] By selecting the $3N - 6$ eigenvectors corresponding to the largest eigenvalues, the determinant of the $\mathbf{G'_m}$ matrix will have its maximum value, i.e., selecting the least redundant set of coordinates. This is the same procedure used for defining delocalized internal coordinated for use in geometry optimizations.[23] Alternatively, the symmetry matrix can be determined using a rule-based combination of primitive coordinates, of which the procedure for defining natural internal coordinates (NIC) is the most well-known.[22,28] The main limitation of NIC is the difficulty in defining nonredundant coordinates for systems containing fused and/or large rings. Furthermore, it may be difficult to extend the principles for constructing NIC to general types of internal coordinates, such as for example long-range coordinates discussed below.

If the PCA extraction of symmetry coordinates is performed for all the primitive coordinates together, each symmetry coordinate will in general contain contributions from all possible types, i.e., a mixture of S, B, and T type coordinates. von Arnim and Ahlrichs have in connection with geometry optimization suggested that the "hard" and "soft" degrees of freedom should be separated.[28] This may be done by partitioning the primitive coordinates into blocks and performing a PCA on each block separately. After extracting the nonredundant combinations from the first block, the primitive coordinates in the second block are orthogonalized against the symmetry coordinates from the first block, before being subjected to a PCA, and so on for all subsequent blocks. In this procedure, there are three choices: how the primitive coordinates are partitioned into blocks, the order by which the blocks are subjected to the PCA, and the threshold on eigenvalues for deciding whether a given PCA component is used for defining a symmetry coordinate. It is advantageous if the block partitioning of coordinates is done in such a way that the PCA for each block results in a clear separation of the eigenvalues of the redundant and nonredundant combinations. If the eigenvalues form a near continuous set, the number of symmetry coordinates extracted from each block will depend on a somewhat arbitrary cutoff parameter, and the number of coordinates selected in a given block will influence the number of coordinates extracted from all subsequent blocks due to the orthogonalization. By partitioning the primitive coordinates into blocks, the $\mathbf{G'_m}$ matrix will always become more singular (smaller determinant), as a block-sequential PCA for extracting symmetry coordinates corresponds to neglecting coupling elements between some of the **b** vectors, but for many

***Table 1.*** Vibrational Analysis for the $C_{60}$ Molecule (60 atoms, 174 vibrational modes)

| | Symmetry coordinates | | | | Normal modes | | |
|---|---|---|---|---|---|---|---|
| PCA order | $N_S$ | $N_B$ | $N_T$ | $-\log(\lvert\mathbf{G'}_m\rvert)$ | $N_S$ | $N_B$ | $N_T$ |
| SBT | 56.9 | 49.4 | 67.7 | 115 | 42.7 | 42.6 | 88.7 |
| S, B, T | 90 | 84 | 0 | 192 | 78.3 | 95.7 | 0 |
| S, T, B | 90 | 0 | 84 | 144 | 73.5 | 0 | 100.5 |
| T, B, S | 29 | 0 | 145 | 202 | 35.2 | 0 | 138.8 |

$N_S$, $N_B$, and $N_T$ indicates the sum of squared coefficients for stretch, bend, and torsion coordinates in the symmetry coordinates and in the normal modes. $\log(\lvert\mathbf{G'_m}\rvert)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion.

reasonable choices it will still be well within limits that allow safe numerical handling.

All results in the present paper have been obtained with the OPLS force field,[39] with force constants calculated by the Tinker program,[40] and vibrational analysis performed with a locally modified version of the Gamess-US program.[41]

## Results

The key concept in the present paper is to investigate how to define primitive coordinates and extract symmetry coordinates to provide a useful set of coordinates for describing molecular vibrations without compromising numerical accuracy. The freedom in defining symmetry coordinates is illustrated in Table 1 for the $C_{60}$ fullerene. A total of 990 primitive coordinates (90 S, 540 B, and 360 T) can be defined for this system, while only 174 nonredundant combinations are possible. If a PCA analysis is performed on all 990 primitive coordinates together, the 174 symmetry coordinates in total contain 56.9 S, 49.4 B, and 67.7 T primitive coordinates, as defined by the sum of squares of the coefficients in the rows of the symmetry matrix in eq 18 (coefficients from the PCA eigenvectors of the $\mathbf{G}$ matrix in eq 22). If the three types of internal coordinates instead are analyzed separately in the order S, B, and T, all 90 primitive stretch coordinates are nonredundant. Once these are defined as symmetry coordinates, there are 84 nonredundant combinations of the bending coordinates, and all torsional coordinates consequently become redundant. If the PCA analysis instead is performed in the order S, T, and B, 84 nonredundant combinations of the torsional coordinates are extracted, and all bending coordinates becomes redundant. The last row in Table 1 shows that if the PCA analysis is done in the order T, B, and S, the nonredundant vibrational space is described by 29 stretching and 145 torsional coordinates. The $\log(\lvert\mathbf{G'_m}\rvert)$ values show that all four of these sets of symmetry coordinates can be handled without numerical problems. The four sets of symmetry coordinates in Table 1 represent limiting cases where all PCA vectors corresponding to numerically nonzero eigenvalues are selected for each block, which for this system corresponds to eigenvalues larger than 0.01, but other combinations of S, B, and T coordinates could also be generated.

The vibrational analysis corresponds to determining the eigenvalues and -vectors of the $\mathbf{G'_m F_d}$ matrix expressed in the selected symmetry coordinates, and the normal modes can subsequently be back-transformed to the primitive $\mathbf{q}$ coordinates by eq 19. It should be stressed that all four sets of coordinates in Table 1 give the same vibrational frequencies, but the description of the corresponding normal modes in terms of primitive coordinates is substantially different. Since the normal modes expressed in the $\mathbf{q}$ coordinates are not orthogonal, the contribution of the different types of coordinates is not necessarily the same as for the symmetry coordinates, and the composition in terms of S, B, and T coordinates is also shown in Table 1. While the nonorthogonality introduces some changes in the importance of each type of coordinate, the results show that there is considerable freedom in selecting coordinates for the vibrational analysis.

One of our main motivations for the present work was to perform vibrational analysis of proteins, where the conformational degrees of freedom are determined primarily by torsional coordinates. In order to simplify the description of the vibrations related to conformational transitions, it is of interest to perform the PCA in the order S, B, and T in order to minimize the number of torsional coordinates. We furthermore propose a separation of the torsional coordinates into four groups, defined as follows:

- Hard (HD) torsions: The two central atoms either form a (partial) double bond or are part of a (small) ring system (only five- and six-membered rings are relevant for proteins). Rotation around amide bonds belongs to this class. For three-coordinated atoms with a near-planar geometry, these torsional coordinates describe the out-of-plane movement of the central atom.
- Back-Bone (BB) torsions: The two central atoms are part of the peptide backbone but are not connected by an amide bond; i.e., these are the Ramachandran $\phi$ and $\psi$ angles.
- Side-Chain (SC) torsions: The two central atoms belong to an amino acid side chain, or one of them is the backbone $C_\alpha$ atom, but the torsion is not of the soft type.
- Soft (SF) torsions: Rotation around single bonds connected to terminal groups in a side-chain. This class includes all torsional angles describing the rotation of $-$OH, $-$SH, $-CH_3$, and $-NH_3$ groups. These four torsional blocks are analyzed in the order HD, BB, SC, and SF when the symmetry coordinates are extracted by PCA. Table 2 shows the vibrational analysis for the angiotensin octapeptide (amino acid sequence DRVYIHPF) in an extended conformation, where 158 S, 281 B, and 411 T primitive coordinates can be generated on the basis of atomic distance criteria. The torsional coordinates can, according to the above classification, be separated into 151 $T_{HD}$, 90 $T_{BB}$, 120 $T_{SC}$, and 50 $T_{SF}$ coordinates. The combined PCA treating all SBT coordinates together provides symmetry coordinates corresponding to 154.7 S, 155.3 B, and 149.0 T primitive coordinates. Performing the PCA on the block-separated coordinates gives the results in the second row in Table 2, where the number of torsional coordinates now is reduced to 56. The $\log(\lvert\mathbf{G'_m}\rvert)$ value, however, shows that the symmetry coordinates between blocks have become somewhat linearly dependent. The main reason for this is a strong coupling between the B and $T_{HD}$ coordinates, as shown by the results in the third row of Table 2, where the $T_{HD}$'s are allowed to mix with the bending coordinates during the PCA. The necessity of allowing the B and $T_{HD}$ coordinates to mix

***Table 2.*** Vibrational Analysis for the Angiotensin Octapeptide (155 atoms, 459 vibrational modes)[a]

| | Symmetry Coordinates | | | | | | | Normal Modes | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PCA order | $N_S$ | $N_B$ | $N_{T\text{-}HD}$ | $N_{T\text{-}BB}$ | $N_{T\text{-}SC}$ | $N_{T\text{-}SF}$ | $-\log(\lvert\mathbf{G'_m}\rvert)$ | $N_S$ | $N_B$ | $N_{T\text{-}HD}$ | $N_{T\text{-}BB}$ | $N_{T\text{-}SC}$ | $N_{T\text{-}SF}$ |
| SBT | 154.7 | 155.3 | 56.0 | 33.3 | 41.4 | 18.6 | 308 | 99.5 | 104.7 | 77.1 | 50.0 | 80.1 | 47.7 |
| S, B, T | 158 | 245 | 19 | 15 | 15 | 7 | 541 | 114.1 | 163.1 | 37.0 | 40.0 | 65.5 | 39.1 |
| S, $BT_{HD}$, T | 158 | 195.7 | 68.3 | 15 | 15 | 7 | 354 | 107.4 | 134.9 | 96.0 | 32.3 | 53.8 | 34.6 |
| NIC | 158 | 205 | 59 | 15 | 15 | 7 | 376 | 111.0 | 148.0 | 67.4 | 36.0 | 59.2 | 37.5 |

[a] $N_S$, $N_B$, $N_{T\text{-}HD}$, $N_{T\text{-}BB}$, $N_{T\text{-}SC}$, and $N_{T\text{-}SF}$ indicate the sum of squared coefficients for stretch, bend, and torsion (hard, back-bone, side-chain, soft) coordinates in the symmetry coordinates and in the normal modes. $\log(\lvert\mathbf{G'_m}\rvert)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion. NIC indicates natural internal coordinates, using 170° as a criterion for using improper torsional coordinates for near-planar tri-coordinated atoms.

is that the HD torsional coordinates describe the out-of-plane motion for near-planar atoms, which are strongly coupled to the corresponding in-plane bending motion. The last row in Table 2 shows the results obtained using natural internal coordinates for the analysis, where an improper torsional coordinate has been used to describe the out-of-plane motion for tricoordinated atoms when the torsional angle was larger than 170°. It is seen that the PCA-based definition of the symmetry matrix gives very similar results to using rule-based NIC but avoids the problems of defining NIC for large rings[42] and using a somewhat arbitrary parameter for deciding when to use out-of-plane coordinates instead of regular bending coordinates.

The normal-mode analysis in terms of contribution from the primitive coordinates is shown in the right-hand part of Table 2. The results in the third row show that the 15 $T_{SC}$ coordinates make contributions corresponding to a total of 53.8 normal modes, and the seven $T_{SF}$ coordinates make contributions to 34.6 modes. The difference between the contribution of the primitive coordinates to the symmetry coordinates and the normal modes is a measure of the nonorthogonality of the normal modes in the primitive coordinates. The significance is that the PCA can extract internal degrees of freedom that span a proportionally larger part of the vibrational space than they span in terms of symmetry coordinates.

The $\log(\lvert\mathbf{G'}_m\rvert)$ value can be used as a criterion for selecting one set of symmetry coordinates over another,[37] but other criteria are also possible. In the normal coordinate system, the $\mathbf{G'_m}^{-1}$, $\mathbf{F_d}$, and $\mathbf{G'_m F_d}$ matrices are diagonal, and symmetry coordinates that make these matrices more diagonal dominant may thus indicate that they are more natural coordinates for describing the vibrational problem. As the diagonal elements of the $\mathbf{G'_m F_d}$ matrix in the normal coordinate system are proportional to the square of the frequencies, we suggest that the difference between $\sqrt{(\mathbf{G'_m F_d})_{ii}}$ and $\nu_i$ over all or a selected range of frequencies can be used as an indicator of whether a given set of symmetry coordinates is better than other alternatives.

Table 3 shows the average diagonal and off-diagonal elements of the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices, as well as the $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference averaged over all frequencies. A complete separation of the S, B, and T primitive coordinates (row 2 in Table 3) leads to a small value of the $\mathbf{G'_m}$ determinant and results in an increase in the $\mathbf{F_d}$ matrix elements by roughly 3 orders of magnitude, which clearly is unfavorable. The coordinate separation where the HD torsions are combined with the bending coordinates during

***Table 3.*** Vibrational Analysis for the Angiotensin Octapeptide (155 atoms, 459 vibrational modes)[a]

| PCA order | $-\log(\lvert\mathbf{G'_m}\rvert)$ | $\langle\lvert F_{ii}\rvert\rangle$ | $\langle\lvert F_{ij}\rvert\rangle$ | $\langle\lvert GF_{ii}\rvert\rangle$ | $\langle\lvert GF_{ij}\rvert\rangle$ | $\langle\lvert\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i\rvert\rangle$ |
|---|---|---|---|---|---|---|
| SBT | 308 | 0.21 | 0.005 | 0.089 | 0.0027 | 471 |
| S, B, T | 541 | 1227.86 | 14.75 | 0.090 | 0.0094 | 220 |
| S, $BT_{HD}$, T | 354 | 0.25 | 0.004 | 0.089 | 0.0013 | 223 |
| NIC | 376 | 0.26 | 0.002 | 0.089 | 0.0003 | 126 |

[a] NIC indicates natural internal coordinates, using 170° as a criterion for using improper torsional coordinates for near-planar tri-coordinated atoms. $\langle\lvert F_{ii}\rvert\rangle$, $\langle\lvert F_{ij}\rvert\rangle$, $\langle\lvert GF_{ii}\rvert\rangle$, and $\langle\lvert GF_{ij}\rvert\rangle$ are average diagonal and off-diagonal elements of the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices, respectively (in atomic units). The last column is the average difference between the square root of the diagonal elements of the $\mathbf{G'_m F_d}$ matrix and the frequencies (in $cm^{-1}$).

the PCA, however, makes both the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices more diagonal dominant, as well as reducing the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference by roughly a factor of 2. A separation of S, B, and T coordinates, where the HD torsions are combined with the bending coordinates for defining the symmetry matrix, thus appears to be preferable over a combined approach where all primitive coordinates are allowed to mix. The use of natural internal coordinates makes the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices even more diagonal and reduces the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference by another a factor of 2.

In large proteins, the low-frequency modes often describe large-scale movements corresponding to elongation, bending, and twisting motions of the whole system or of large domains relative to each other. In either Cartesian or internal coordinates of the type discussed above, these normal modes are described as a linear combination of many coordinates, each of which enters the normal mode with a small coefficient. The existence of such large-scale movements suggests that the primitive set of coordinates should be augmented with long-range coordinates to allow a more compact description of the low-frequency modes. One possibility is to define long-range stretch, bend, and torsion coordinates between $C_\alpha$ atoms of each residue, and such coordinates can be generated automatically, analogous to the corresponding short-range coordinates. We have considered a scheme based on a list of $C_\alpha$ atoms ordered according to the protein backbone connectivity, as illustrated in Figure 1, where long-range coordinates are defined between $C_\alpha$ atoms that are separated by a fixed number of residues, denoted stride. A stride of 1 means that long-range coordinates are defined between all $C_\alpha$ atoms in neighboring residues (1−2, 2−3, 3−4, etc.); a stride of 2 means that long-range coordinates are defined between $C_\alpha$ atoms in every

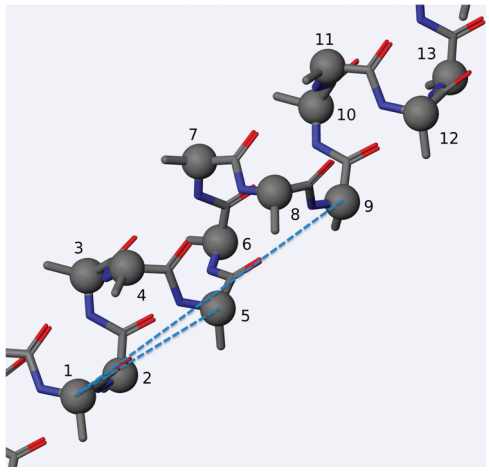



***Figure 1.*** Illustration of a helical conformation of a polyalanine peptide with $C_\alpha$ atoms highlighted and hydrogen atoms omitted. Long-range stretching coordinates with a stride of 4 and 8 are indicated by dashed lines.

***Table 4.*** The Effect of Adding Long-Range Coordinates on Vibrational Analyses for Helical and Extended Conformations of a 50 Residue Polyalanine Peptide[a]

| | Helix | | | Extended | | |
|---|---|---|---|---|---|---|
| stride | $S_{LR}$ | $B_{LR}$ | $T_{LR}$ | $S_{LR}$ | $B_{LR}$ | $T_{LR}$ |
| none | | 34.6 | | | 36.2 | |
| 1 | 32.8 | 25.9 | 29.6 | 48.8 | 16.7 | 5.1 |
| 2 | 5.6 | 12.5 | 9.8 | 0.7 | 17.3 | 4.2 |
| 3 | 4.6 | 26.1 | 19.4 | 5.0 | 17.5 | 4.8 |
| 4 | 1.0 | 14.2 | 19.5 | 0.3 | 15.7 | 3.8 |
| 5 | 2.4 | 13.4 | 16.4 | 4.5 | 14.5 | 4.8 |
| 6 | 1.1 | 15.4 | 12.8 | 0.7 | 14.7 | 4.0 |
| 7 | 1.6 | 14.9 | *b* | 5.1 | 10.0 | *b* |
| 8 | 0.6 | *b* | *b* | 1.3 | *b* | *b* |
| 9 | 1.2 | *b* | *b* | 4.4 | *b* | *b* |

[a] Table entries show $\langle|\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i|\rangle$ for the lowest five vibrational frequencies (in $cm^{-1}$). $S_{LR}$, $B_{LR}$, and $T_{LR}$ indicate long-range stretch, bend, and torsional coordinates, respectively. [b] Less than five long-range coordinates added.

second residue (1−3, 3−5, 5−7, etc.); a stride of 3 means that long-range coordinates are defined between $C_\alpha$ atoms in every third residue (1−4, 4−7, 7−11, etc.); etc. The usefulness of stride restrictions is perhaps best illustrated by considering the α helix in Figure 1, where a stride of 4 corresponds to linking $C_\alpha$ atoms between residues having made approximately a full helix turn. Long-range coordinates with strides of 4, 8, etc. may thus be naturally well-suited for describing deformations of α-helix moieties. In the PCA extraction of symmetry coordinates, the long-range coordinates of a given type are collected in a separate block and are always treated before the corresponding short-range coordinates; i.e., the ordering is $S_{LR}$, S, $B_{LR}$, $B+T_{HD}$, $T_{LR}$, $T_{BB}$, $T_{SC}$, and $T_{SF}$. The PCA ensures that the nonredundant long-range coordinates are selected first, and linear combinations of short-range coordinates that span the same space are removed.

Table 4 shows the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference for the five lowest modes for helical and extended conformations of a 50 residue polyalanine peptide as a function of adding long-range coordinates with increasing strides. For the helical conformation, the five lowest vibrational frequencies are 2.16, 2.16, 5.63, 5.67, and 6.25 $cm^{-1}$, while the corresponding values for the extended conformation are 0.14, 0.15, 0.34, 0.47, and 0.81 $cm^{-1}$, and these describe global deformation modes. In the absence of long-range coordinates, the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ deviations are 34.6 and 36.2 $cm^{-1}$, respectively, while the corresponding values using natural internal coordinates are 45.1 and 62.9 $cm^{-1}$ (not shown in Table 4). The addition of long-range stretching coordinates with strides larger than 1 markedly reduces the deviation, while long-range bending or torsional coordinates are less effective. For the extended conformation, the addition of long-range stretching coordinates with even strides has a larger effect than with odd strides. We have checked that a similar behavior is observed for a 49 residue polyalanine; i.e., the alternating effect is not related to the specific length of the polypeptide. For both conformations, a stride of 4 appears to be especially effective in reducing the deviation.

Table 5 shows the effects on the symmetry coordinates and normal modes by the addition of long-range stretch, bend, and torsion coordinates with a stride of 4 for helical and extended conformations of a 50 residue polyalanine peptide. For the symmetry coordinates, the long-range coordinates simply replace a corresponding number of backbone torsional coordinates, and the $\log(|\mathbf{G'_m}|)$ value changes only slightly. The main effect on the description of

***Table 5.*** The Effect of Adding Long-Range Stretching Coordinates with a Stride of 4 on the Vibrational Analyses for Helical and Extended Conformations of a 50 Residue Polyalanine Peptide[a]

| | symmetry coordinates | | | | | normal modes | | average 5 mode % composition | | |
|---|---|---|---|---|---|---|---|---|---|---|
| conformation | $N_{S\text{-}LR}$ | $N_{B\text{-}LR}$ | $N_{T\text{-}LR}$ | $N_{T\text{-}BB}$ | $-\log(\lvert\mathbf{G'}_m\rvert)$ | $N_{X\text{-}LR}$ | $N_{T\text{-}BB}$ | $C_{X\text{-}LR}$ | $C_{T\text{-}BB}$ | $C_{T\text{-}other}$ |
| helix | 0 | 0 | 0 | 96 | 1239 | 0 | 241.8 | 0 | 63.9 | 33.3 |
| | 11 | 0 | 0 | 85 | 1225 | 24.7 | 220.2 | 54.9 | 26.9 | 17.1 |
| | 0 | 10 | 0 | 86 | 1244 | 1.5 | 231.3 | 8.0 | 55.4 | 33.8 |
| | 0 | 0 | 9 | 87 | 1226 | 51.7 | 209.0 | 65.4 | 19.9 | 13.9 |
| extended | 0 | 0 | 0 | 96 | 1235 | 0 | 270.7 | 0 | 98.5 | 1.3 |
| | 11 | 0 | 0 | 85 | 1246 | 19.7 | 244.9 | 17.5 | 80.3 | 1.9 |
| | 0 | 10 | 0 | 86 | 1250 | 4.6 | 252.4 | 33.6 | 60.0 | 5.5 |
| | 0 | 0 | 9 | 87 | 1207 | 254.9 | 148.2 | 100.0 | 0.0 | 0.0 |

[a] $N_{S\text{-}LR}$, $N_{B\text{-}LR}$, $N_{T\text{-}LR}$, and $N_{T\text{-}BB}$ indicate the number of long-range stretch, bend, torsion, and back-bone torsional coordinates, respectively, and $N_{X\text{-}LR}$ indicates the number of the particular type of long-range coordinate. $\log(|\mathbf{G'_m}|)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion. $C_{X\text{-}LR}$, $C_{T\text{-}BB}$, and $C_{T\text{-}other}$ indicate the average composition of the five lowest normal modes in terms of long-range, back-bone torsion, and torsion coordinates other than back-bone, respectively. The total number of normal modes is 1470. The major change by the addition of long-range coordinates is in the number of backbone torsional coordinates and torsional normal modes, and only those are listed.

the normal modes by adding long-range coordinates is also a replacement of the backbone torsional coordinates, but there is a significant difference between the type of long-range coordinate and the helical/extended conformation.

For the helical conformation, long-range stretching and torsional coordinates are more important than long-range bending coordinates. The 11 long-range stretching coordinates contribute to a total of 24.7 normal modes, as measured by the sum of squared coefficients in eq 16. The 10 long-range bending coordinates, on the other hand, only contribute to 1.5 normal modes, while the nine long-range torsional coordinates contribute to 51.7 normal modes. The long-range coordinates are primarily involved in the description of the low frequency normal modes, as illustrated by the average composition of the five lowest normal modes shown in the last three columns in Table 5. In the absence of long-range coordinates, these global deformation modes are described primarily as a combination of backbone torsional coordinates, where the weight of any primitive backbone torsional coordinate to a given normal mode is less than 1%. The addition of only 11 long-range stretching coordinates to the set of 2589 short-range primitive coordinates results in these five modes being described as a combination of long-range stretching and short-range torsional coordinates. The 11 long-range stretching coordinates describe more than 50% of the five lowest normal modes on average, and each individual long-range stretching coordinate accounts for up to ∼25% of a given normal mode. A similar effect is observed for the long-range torsional coordinates.

The extended conformation has a large number of low frequency bending and twisting modes (145 modes below 100 $cm^{-1}$), of which many resemble standing waves for a string. Long-range torsional coordinates are very efficient at describing these modes, and the nine long-range torsional coordinates make contributions corresponding to a total of 254.9 normal modes! The description of the five lowest normal modes accordingly changes from being a linear combination of many backbone torsional coordinates to a linear combination of essentially only the nine long-range torsional coordinates.

Consecutive helical or sheet moieties with ∼50 residues are unusual in actual proteins, and the results in Tables 4 and 5 thus primarily serve as an illustration that long-range coordinates may be advantageous for a compact description of low-frequency large-scale motions in proteins. Instead of representing the low-frequency normal modes as a linear combination of a large number of short-range primitive coordinates, these modes can instead be represented as a combination of a few long-range coordinates with large weights for describing the overall motion and number of short-range coordinates with small weights for describing the local displacements. For globular proteins, where residues may be spatially close without being close in terms of backbone bonding, additional long-range coordinates defined by distance or hydrogen bonding criteria may be advantageous. Which long-range or interstrand coordinates are most advantageous will most likely depend on the nature of the given protein. Vibrational analyses for a number of real proteins are required to probe these effects.

## Summary

The present work shows that there is considerable freedom in choosing internal coordinates and in extracting nonredundant combinations of these for describing vibrational normal coordinates. For large molecular systems, like proteins, the addition of selected long-range coordinates can provide a compact description of especially the low-frequency normal modes. Vibrational normal modes calculated in general coordinates may be of use for analysis purposes, for example, for identifying possible large-scale transformations,[7] for biasing molecular dynamics simulations,[10] or for running dynamics simulations in selected internal coordinates.[43] The use of long-range coordinates may also be of interest in geometry optimizations of large flexible systems.

**Acknowledgment.** This work was supported by grants from the Danish Center for Scientific Computation, the Danish Natural Science Research Council, and the Villum Kann Rasmussen foundation.

## References

(1) Bowman, J. M.; Christoffel, K.; Tobin, F. *J. Chem. Phys.* **1979**, *83*, 905. Heislbetz, S.; Rauhut, G. *J. Chem. Phys.* **2010**, *132*, 124129. Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2140. Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149.

(2) Miloshevsky, G. V.; Jordan, P. C. *Structure* **2006**, *14*, 1241. Miloshevsky, G. V.; Jordan, P. C. *Structure* **2007**, *15*, 1654. Miloshevsky, G. V.; Hassanein, A.; Jordan, P. C. *J. Mol. Struct.* **2010**, *972*, 1.

(3) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. *Chem. Rev.* **2010**, *110*, 1463.

(4) Kumar, P.; Joshi, D. C.; Akif, M.; Akhter, Y.; Hasnain, S. E.; Mande, S. C. *Biophys. J.* **2010**, *98*, 305.

(5) Ma, J. *Structure* **2005**, *13*, 373.

(6) Thomas, A.; Field, M. J.; Perahia, D. *J. Mol. Biol.* **1996**, *261*, 490.

(7) Tama, F.; Sanejouand, Y.-H. *Protein Eng.* **2001**, *14*, 1.

(8) Dobbins, S. E.; Lesk, V. I.; Sternberg, J. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10390.

(9) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. *Annu. Rev. Biophys.* **2010**, *39*, 23.

(10) Isin, B.; Schulten, K.; Tajkhorshid, E.; Bahar, I. *Biophys. J.* **2008**, *95*, 789.

(11) Ma, J. *Structure* **2005**, *13*, 373. Bahar, I.; Rader, A. *J. Curr. Opin. Struct. Biol.* **2005**, *15*, 586.

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

(13) Njegic, B.; Gordon, M. S. *J. Chem. Phys.* **2008**, *129*, 164107.

(14) Brüschweiler, R. *J. Chem. Phys.* **1995**, *102*, 3396. Brooks, B. R.; Janezic, D.; Karplus, M. *J. Comput. Chem.* **1995**, *16*, 1522.

(15) During the course of this work, McIntosh published an essentially identical procedure for determining vibrational frequencies in general coordinates: McIntosh, D. F. *Theor. Chem. Acc.* **2010**, *125*, 177.

(16) Manson, S. A.; Law, M. M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2848. Venderell, O.; Gatti, F.; Lauvergnat, D.; Mayer, H.-D. *J. Chem. Phys.* **2007**, *127*, 184302. Stare, J.; Balint-Kurti, G. G. *J. Phys. Chem. A* **2003**, *107*, 7204.

(17) Kamiya, K.; Sugawara, Y.; Umeyama, H. *J. Comput. Chem.* **2003**, *24*, 826.

(18) Boatz, J. A.; Gordon, M. S. *J. Phys. Chem.* **1989**, *93*, 1819.

(19) Konkoli, Z.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 1. Konkoli, Z.; Larsson, J. A.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 11. Konkoli, Z.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 29. Konkoli, Z.; Larsson, J. A.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 41.

(20) Wilson, E. B., Jr. *J. Chem. Phys.* **1939**, *7*, 1047. Wilson, E. B., Jr. *J. Chem. Phys.* **1941**, *9*, 76.

(21) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill Book Company, Inc.: New York, 1955.

(22) Fogarasi, G.; Zhou, X.; Taylor, P. W.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191.

(23) Baker, J.; Kessi, A.; Delley, B. *J. Chem. Phys.* **1996**, *105*, 192.

(24) Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.

(25) Bakken, V.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 9160.

(26) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.

(27) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.

(28) von Arnim, M.; Ahlrichs, R. *J. Chem. Phys.* **1999**, *111*, 9183.

(29) Levitt, M.; Sander, C.; Stern, P. S. *J. Mol. Biol.* **1985**, *181*, 423. Alexandrov, V.; Smith, D. M. A.; Rostkowska, H.; Nowak, M. J.; Adamowicz, L.; McCarthy, W. *J. Chem. Phys.* **1998**, *108*, 9685.

(30) Baker, J.; Puley, P. *J. Chem. Phys.* **1996**, *105*, 11100. Baker, J.; Puley, P. *J. Comput. Chem.* **2000**, *21*, 69. Swart, M.; Bickelhaupt, F. M.*Int. J. Quantum Chem.* **2006**, *106*, 2536. Maslen, P. E. *J. Chem. Phys.* **2005**, *122*, 014104.

(31) Stare, J. *J. Chem. Inf. Model.* **2007**, *47*, 840.

(32) Stare, J.; Balint-Kurti, G. G. *J. Phys. Chem. A* **2003**, *107*, 7204.

(33) Eckart, C. *Phys. Rev.* **1935**, *47*, 552.

(34) Miyazawa, T. *J. Chem. Phys.* **1958**, *29*, 246.

(35) The standard vibrational analysis in Cartesian coordinates corresponds to the **B** matrix being a unit matrix, and thus $\mathbf{G}_m = \mathbf{M}^{-1}$.

(36) Baker, J.; Kinghorn, D.; Pulay, P. *J. Chem. Phys.* **1999**, *110*, 4986.

(37) Németh, K.; Challacombe, M.; van Veenendaal, M. *J. Comput. Chem.* **2010**, *31*, 2078.

(38) A force constant weighting of the **B** matrix elements has been proposed by Lindh, R.; Bernhardsson, A.; Schütz, M. *Chem. Phys. Lett.* **1999**, *303*, 567. This weighting corresponds to a partial least-squares (PLS) approach for defining the symmetry matrix instead of a PCA.

(39) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(40) Tinker v5.1. http://dasher.wustl.edu/tinker/ (accessed 1/1/2010).

(41) 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.

(42) Large rings in proteins are formed by disulfide bridges between cysteine residues.

(43) Pulay, P.; Paizs, B. *Chem. Phys. Lett.* **2002**, *353*, 400. Jaqaman, K.; Ortoleva, P. J. *J. Comput. Chem.* **2002**, *23*, 484. Miao, Y.; Ortoleva, P. J. *J. Comput. Chem.* **2008**, *30*, 423. Sherif, Z.; Ortoleva, P. *J. Stat. Phys.* **2008**, *130*, 669. Lee, S.-H.; Palmo, K.; Krimm, S. *Comput. Chem.* **2007**, *28*, 1107. Henriksson, K. O.; Pesonen, J. *J. Comput. Chem.* **2010**, *31*, 1882.

CT100463A

JCTC Journal of Chemical Theory and Computation

# Improved Replica Exchange Method for Native-State Protein Sampling

Samuel L. C. Moors,* Servaas Michielssens, and Arnout Ceulemans

*Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, 3001 Heverlee, Belgium*

Received August 29, 2010

**Abstract:** We present a new replica exchange method, designed for optimal native state protein sampling in explicit solvent, called replica exchange with flexible tempering (REFT). The method was built upon the recently introduced replica exchange with solute tempering (REST). The potential function is adapted to direct the conformational search toward interdomain movements and the flexible portions of the protein. We demonstrate the improved sampling efficiency of REFT compared to the original REST for the bacteriophage T4 lysozyme.

## Introduction

The replica exchange method (REM) is a parallel simulation method aimed to improve the sampling efficiency of complex systems with rugged energy landscapes such as proteins.[1−3] In REM, multiple replicas are simulated in parallel in different thermodynamic states. At regular intervals during the simulation, exchanges between replicas are attempted. The acceptance probability for exchange is based on the detailed balance condition, which ensures that the canonical ensemble is conserved in the limit of an infinitely long simulation time.[4] In its original implementation, the temperature was chosen as the state variable. Exchanges between low and high-temperature states allow the system to cross high barriers much more easily. The temperature REM (T-REM) method proved to be very powerful for small systems such as peptides and small proteins.[5−7] However, as the number of replicas increases with the square root of the system degrees of freedom,[8] its use for medium sized proteins or larger becomes impractical.

To overcome the limitations of the original T-REM, many variations have been suggested in the literature.[9] Hamiltonian REM (H-REM), which uses the potential energy function as the state variable, is particularly promising.[8] In H-REM only parts of the Hamiltonian are scaled between replicas[10] as opposed to the temperature, which is equivalent to a uniform scaling of the potential energy. Reported H-REM implementations include scaled hydrophobicity and hydrophobic aided REM (mimicking the chaperone effect by lowering the hydrophobic interactions),[8,11] phantom chain and soft-core REM (allowing partial atomic overlaps),[8,12] REM with peptide backbone biasing potential (lowering the backbone dihedral angle rotational barriers).[13] Zacharias developed a H-REM method that applies a penalty potential to drive the protein along the soft modes, which were determined from elastic network model calculations.[14]

Berne and co-workers introduced an interesting REM variant, called replica exchange with solute tempering (REST).[15] In REST, the potential energy scales with the temperature in such a way that the solvent appears cold even at high temperatures. The mutual interaction energy between water molecules disappears from the exchange acceptance criterion, giving rise to higher exchange probabilities. Because the bulk of the system degrees of freedom stems from the water molecules, far less replicas are needed.

In this article, we propose to advance the REST method for native-state sampling by using a dual scale approach. Proteins can be subdivided into rigid and flexible regions. Rigid domains usually consist of (single or multiple) secondary structure elements like α-helices and $\beta$-sheets. Rigid domains are often connected by flexible hinges or loops, allowing large movements between them. We hypothesize that the sampling rate of native proteins is limited by the slow large-scale movements between rigid domains. In the new method, we selectively heat-up the flexible parts, keeping the rigid domain interactions cold. As such, interdomain motions are facilitated, while the domains themselves maintain a nativelike state. We refer to this method as replica

* Corresponding author phone: (32)16/32.73.84; fax: (32)16/32.79.92; e-mail: sam.moors@chem.kuleuven.be.

10.1021/ct100493v 
Published on Web 12/07/2010

exchange with flexible tempering (REFT). The REFT method was applied to the native-state sampling of the bacteriophage T4 lysozyme (T4L) and the sampling efficiency was compared with the REST method.

## Materials and Methods

**Replica Exchange with Solute Tempering.** The REST method is a combination of H-REM and T-REM. In REST, the potential energy function $E_0$ is decomposed into three terms

$$E_0 = E_{pp} + E_{ww} + E_{pw} \quad (1)$$

where pp, ww, and pw correspond to the internal protein−protein, water−water, and protein−water interactions respectively. The ww and pw energy terms are scaled with the simulation temperature

$$E_i = E_{pp} + (\beta_0/\beta_i)E_{ww} + \sqrt{\beta_0/\beta_i}E_{pw} \quad (2)$$

where $\beta$ is the reciprocal temperature $1/k_BT$ with $k_B$ the Boltzmann constant. The indices 0 and $i$ refer to the lowest and $i$th temperature replica. Note that we have slightly adapted the original scaling factor of the pw term $(\beta_0 + \beta_i)/2\beta_i$,[15] to allow defining the rigid and flexible atoms at the force field level.

**Replica Exchange with Flexible Tempering.** The REFT method further categorizes the protein atoms as either rigid or flexible. The terms $E_{pp}$ and $E_{pw}$ are further subdivided

$$E_0 = E_{ff} + E_{rr} + E_{ww} + E_{fr} + E_{fw} + E_{rw} \quad (3)$$

where ff, rr, ww, fr, fw, and rw represent the flexible−flexible, rigid−rigid, water−water, flexible−rigid, flexible−water, and rigid−water interactions, respectively. The following scaling was used for the potential energy of the higher-temperature replicas

$$E_i = E_{ff} + (\beta_0/\beta_i)E_{ww,rr,wr} + \sqrt{\beta_0/\beta_i}E_{fw,fr} \quad (4)$$

As in eq 2, this scaled potential allows defining the rigid and flexible atoms on the force field level by appropriately scaling the parameters of individual atoms, pairs, triplets and quadruplets. Both REST and REFT satisfy the detailed balance condition, and thus lead to the canonical ensemble in the limit of infinite sampling.

**Determination of Rigid Domains.** The floppy inclusion and rigid substructure topography (FIRST) program identifies rigid domains in a single protein conformation based on geometrical constraints to model covalent bonds, hydrogen bonds, and hydrophobic tethers.[16] Constraints are defined based on a cutoff energy of interaction. The FIRST analysis was performed on the closed-state X-ray structure (2LZM). Only clusters with more than 15 atoms were included. The constraint cutoff energy value was chosen such that the number of rigid atoms is approximately one-third of all protein atoms. For T4L, a cutoff energy of 0.5 kcal/mol was used, resulting in 12 rigid domains (Figure 1).

**Molecular Dynamics Simulations.** Molecular dynamics (MD) simulations were performed with the AMBER ff03 force field[17] using the open-source *GROMACS 4.0* software package, which was adapted to account for exchanges between different Hamiltonian states.[18] Long-range electrostatics were computed with the particle mesh Ewald method.[19] The nonbonded cutoff was set to 1.2 nm. A Fourier spacing of 0.19 nm was used. All covalent bonds were constrained using P-LINCS.[20] The integration time step was 2 fs. The neighbor list was updated every 10 steps. The system temperature was controlled by the velocity rescaling thermostat, which preserves the canonical ensemble.[21] Exchanges between replicas were attempted every 1 ps.

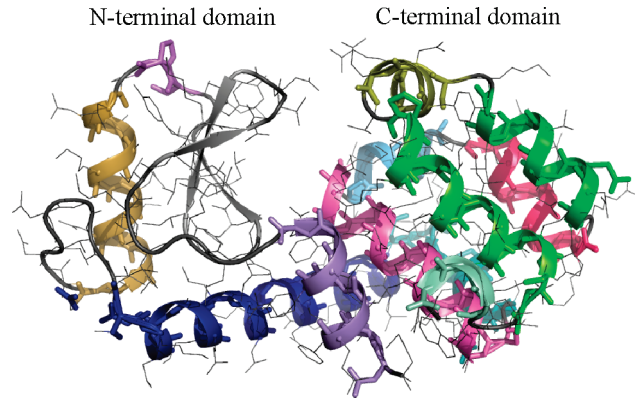



***Figure 1.*** The T4L rigid domains as determined by FIRST. Each colored region represents a different rigid domain. The gray-colored side chains, $\beta$-sheet and loops are flexible. The dark blue α-helix connects the N-terminal and C-terminal domains.

The initial T4L structure was obtained from the X-ray structure (2LZM). Protonation states were determined using the H++ web server.[22] The charged protein was neutralized with $Cl^-$ ions and solvated in a dodecahedral box using 9093 TIP3P water molecules. The neutral solvated system was relaxed, heated to 300 K using position restraints and briefly equilibrated for 50 ps. Next, one REST (s1) and one REFT (f1) simulation were started using identical starting structures for each replica. After 2 ns of conventional MD, again one REST (s2) and one REFT (f2) simulation were started. The REM simulations were carried out for 120 ns, the last 60 ns of which were used for analysis. The f1 simulation was continued until 360 ns for further analysis of the equilibrated ensemble. Replica temperatures were 300.0, 308.4, 317.4, 326.8, 336.6, 346.8, 357.3, 368.2, 379.5, 391.2, 403.3, 415.8, 428.7, 442.1, 455.9, 470.2, 484.9, 500.1, 515.9, 532.2, 549.0, 566.3, 584.3, 602.8 K for REFT, and 300.0, 307.8, 315.9, 324.1, 332.6, 341.2, 350.2, 359.3, 368.7, 378.3, 388.2, 398.3, 408.7, 419.4, 430.3, 441.5, 453.1, 464.9, 477.0, 489.5, 502.2, 515.4, 528.8, 542.6, 556.8, 571.3, 586.2, 601.5 K for REST. The replica temperatures were adapted from an exponentially distributed temperature set, based on preliminary 2 ns replica exchange simulations. Special care was taken to obtain pairwise exchange probabilities between all neighboring replicas close to 20%.

The cleft volume was calculated by generating a spherical grid of oxygen atoms with a radius of 1.1 nm around the center of the $C^\alpha$ atoms of residues Glu11, Gly30, and Phe104. These three residues are located at the active site core inside the cleft. The grid spacing was 0.1 nm. Grid atoms overlapping with protein atoms were removed and the

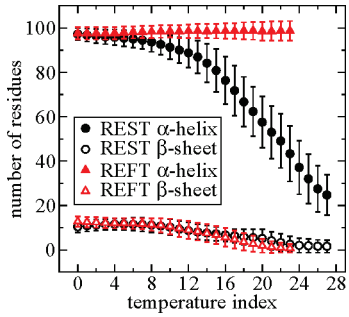



***Figure 2.*** Number of residues that are part of an α-helix or β-sheet as a function of temperature index, averaged over the last 60 ns of the 120 ns simulation. The temperature indices denote the serial numbers of the consecutive thermodynamic states, starting with zero for the 300 K state. Error bars represent the standard deviations of the averages.

number of remaining grid atoms was counted. The principal component analysis (PCA) was carried out on the backbone atoms of a series of 38 X-ray structures as described by de Groot et al.[23] Snapshots of the MD simulations were projected onto the calculated eigenvectors.

## Results

T4L is a 164-residue protein which is known to display large conformational changes upon substrate binding through hinge bending motion.[24,25] The protein consists of two domains connected by a central α-helix. The active site is located in the cleft between the N-terminal and C-terminal domains. The hinge bending motion mediates a closing and opening of the active-site cleft. Starting from the closed state, the equilibration rates of the open and closed states and their relative populations were used to compare the sampling efficiency of the REST and REFT methods.

Four explicit solvent 120 ns simulations were carried out: two independent REST (s1, s2) and two independent REFT (f1, f2) simulations. To span a 300−600 K range with 20% acceptance probability between successive replicas, using REST and REFT 28 and 24 replicas were needed respectively. For comparison, with T-REM we estimate that about 124 replicas would have been needed to span the same exchange rates and temperature range.

We used the FIRST method to identify the rigid domains.[16] The effect of increased rigid domain interactions on the secondary structure is readily seen in Figure 2. In the REST simulation, at high temperatures both the α-helix and β-sheet content decays. Using REFT, α-helix content stays intact even at 600 K. The β-sheet content however decreases with temperature. The β-strands were not identified by FIRST as part of a rigid domain, reflecting the more flexible nature of β-sheets compared to helices.

We define the crossing rate as the number of replicas that cross temperature space between 300 K and a given temperature >300 K. High crossing rates are essential for fast equilibration at low temperature. Compared to the REST simulation, on account of the preservation of secondary structure (Figure 2), the REFT crossing rates have drastically increased (part A of Figure 3). The crossing rate differences become more pronounced as the simulation progresses.

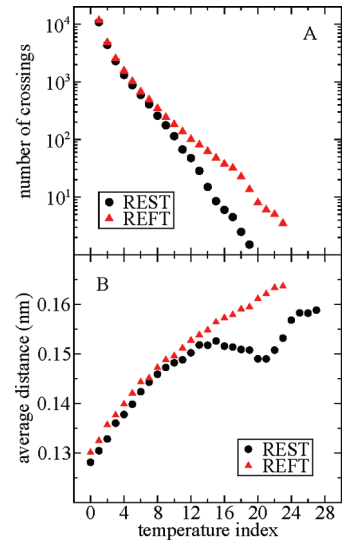



***Figure 3.*** Mobility in temperature and conformational space. (A) Number of temperature crossings between 300 K and the temperature index given in the abscissa. (B) Average distance covered per ps along the first eigenvector of the PCA analysis. Averages were taken over the last 60 ns of the two REST and REFT simulations, respectively.

Above 500 K no single crossing was observed for any of the 28 REST replicas during the last 60 ns. Thus, an effective decoupling occurs between the high and low temperature replicas.

Compared to REST, a larger part of the REFT system is kept cold, which might affect the dynamics in the high-temperature simulations. It is therefore important to evaluate the mobility of the protein in conformational space. The average distance covered every ps along the first PCA eigenvector (Materials and Methods) as a function of temperature is plotted in part B of Figure 3. This collective coordinate corresponds to the hinge bending opening-closing motion. At low temperatures, the mobility along the hinge bending coordinate is comparable. As the temperature increases, part of the kinetic energy is channeled toward the slow hinge bending mode. Above 420 K however, the REST hinge bending mobility starts to degrade. This is easily explained by the loss of native structure, which leads to a distribution of kinetic energy into many local non-native high-frequency motions. By contrast, in the REFT simulation, a steady increase of the hinge bending mobility is seen along the full temperature range.

To better understand how the potential function affects conformational sampling at high temperature, we performed two MD simulations of the REST and REFT states at temperature index 17, but without replica exchanges (Figure 4). The thermodynamic state at temperature index 17 induces strongly elevated protein dynamics, while the crossing rates are still considerable for both methods (parts A and B of Figure 3). Initially good open−closed transition rates are seen for REST-state MD. However, after 4.3 ns the protein starts to unfold, as can be seen from the backbone root-mean-square deviation (bRMSD) plot (part C of Figure 4). As a reference, at 300 K the bRMSD from the X-ray structure fluctuates between 0.07 and 0.47 nm. The unfolding coincides with a sharp decline of the open−closed transitions to

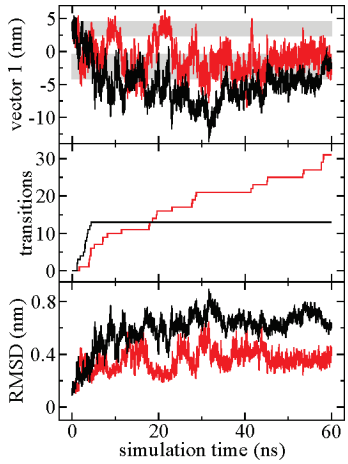



***Figure 4.*** Time evolution of the first PCA eigenvector (A), the cumulative number of transitions between the open and closed states (B), and the bRMSD from the X-ray structure (C) of two high-temperature MD simulations using a potential function corresponding to temperature index 17, which is 464.9 K for REST (black) and 500.1 K for REFT (red). The upper and lower gray bars in (A) indicate the closed and open states, respectively.

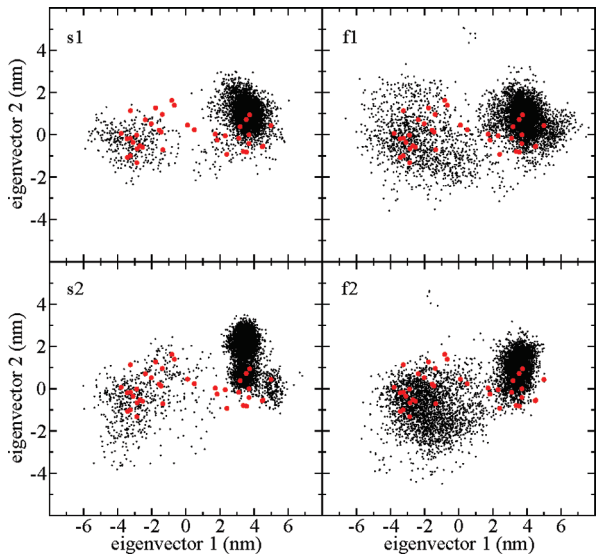



***Figure 5.*** 2D projection of the 300 K trajectories onto the first two eigenvalues of a PCA on the backbone of a series of X-ray structures. REST (s1, s2), REFT (f1, f2). Black dots indicate samples taken every 10 ps. Red circles indicate 38 X-ray reference structures.

zero (part B of Figure 4). For a large part of the simulation time, the system samples a region in conformational space that is non-native (PCA eigenvector 1 < −5, part A of Figure 4). In contrast, the REFT-state MD stays mainly within the native bRMSD region, and the number of open−closed transitions increases steadily throughout the trajectory. For comparison, a 184 ns conventional MD simulation started from the closed state did not yield a single transition to the open state.[26]

In Figure 5, the sampling in the space of the two largest PCA eigenvectors, as well as a series of 38 X-ray structures[23] is shown. The second eigenvector describes a twisting mode. Using REFT, the area of open conformations (eigenvector 1 < 1.25 nm) is populated more densely. At this stage the systems are not yet equilibrated, which explains the differences between the four simulations in Figure 5. All X-ray conformations are sampled by both methods.

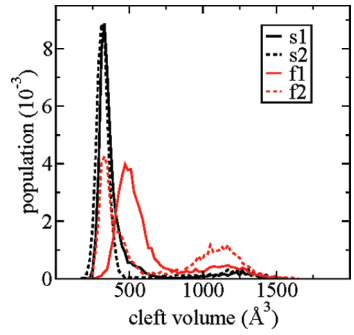



***Figure 6.*** Cleft volume distribution of the two independent REST (s1, s2) and REFT (f1, f2) simulations, from the last 60 ns of the 120 ns simulations at 300 K.

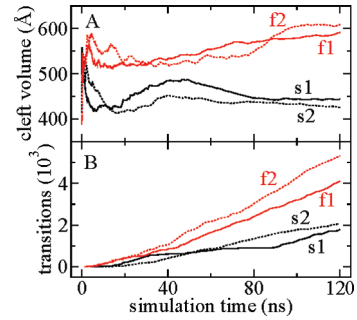



***Figure 7.*** Time evolution of the cumulative average cleft volume (A) and the total number of open−closed transitions (B) in the 300 K state. Starting cleft volumes were 489 Å$^3$ (s1, f1) and 406 Å$^3$ (s2, f2).

Whereas the principal eigenvectors of a PCA analysis represent the collective coordinate with the largest spatial displacement, the cleft volume (Materials and Methods) provides information about the local dynamics at the active site. Again, the REFT simulations display a higher population for the open state (>750 Å$^3$) compared to the REST simulation (Figure 6). Moreover, the closed state distribution is much more dispersed, indicating better local sampling of the active site cleft.

Part A of Figure 7 shows the time evolution of the average cleft volume. Shortly after an initial increase, the REST simulations' cleft volumes decline sharply. A possible explanation for this decline might be that open structures unfold faster than closed structures, and hence are trapped inside the high-temperature states more easily. A similar but much smaller cleft volume decline is also visible in the REFT simulations. After about 10−30 ns, the REFT cleft volumes increase steadily, whereas the REST cleft volumes remain at a much lower level.

The rate of open−closed transitions in the 300 K state, which reflects exchanges between replicas with open and closed conformations, provides an additional measure of the sampling efficiency. As shown in part B of Figure 7, the transition rate is relatively constant for all four simulations, indicating good local replica exchange frequencies. During the last 60 ns of the simulation, the average REFT transition

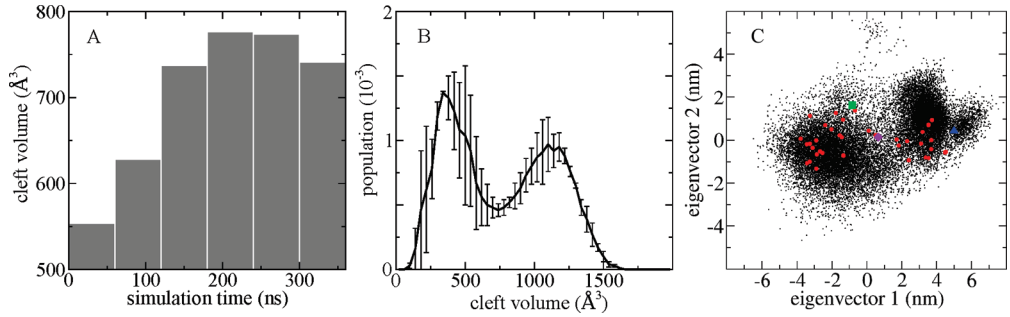



***Figure 8.*** Native state equilibration of T4L in the 300 K state. (A) Evolution of the cleft volume shown as block averages over 60 ns. (B) Cleft volume distribution taken over the last 240 ns of a 360 ns simulation. Error bars represent the standard deviations from the average cleft populations of two 120 ns blocks. The open state has a population of 49 ± 5%. (C) 2D projection of a PCA analysis of the last 240 ns of a 360 ns REFT simulation. The purple diamond indicates the average position in 2D space. The green square and blue triangle represent 150L(c) and 152L, respectively.

rate is 47 transitions/ns, 2.5 times the REST transition rate of 18 transitions/ns. Considering acceptance probabilities ∼20% with one replica exchange attempt every picosecond, the fraction of REFT and REST replica exchanges accompanied by an open−closed transition are ∼23% and ∼9%, respectively.

To estimate the simulation time necessary to obtain a fully equilibrated ensemble, and to get an estimate of the equilibrium distribution, the f1 REFT simulation was extended to 360 ns. Judged by the evolution of the cleft volume (part A of Figure 8), about 120 ns of REFT simulation time seems necessary to reach an equilibrated state. Note that the same starting conformation was used for each replica within a given simulation to avoid any bias in the comparison of both methods. If a diverse set of starting conformations was used across all replicas, a much shorter equilibration time would likely be obtained.

As shown in part B of Figure 8, T4L forms a distinct two-state open−closed equilibrium in solution. Both states occupy approximately half of the population. This two-state model is confirmed by the PCA (part C of Figure 8). In addition, a small third superclosed state is visible in part C of Figure 8, which has a population of ∼5% (eigenvector 1 > 4.9 nm). Here the N- and C-terminal domains are tightly held together by two side chain H-bonds (Asp20 - Gln141) and (Glu22 - Arg137). The latter H-bond is much less prominent, and the former H-bond is virtually nonexistent in the normal closed state.

## Discussion

The efficiency of replica exchange methods depends critically upon (i) the mobility in temperature space and (ii) the mobility in conformational space in the high-temperature range. As was observed by the Berne group, the REST method did not perform well for larger systems.[27] Despite the high acceptance ratios, the rate at which replicas traveled temperature space between the highest and lowest temperature state remained low, effectively decoupling the high and low temperature replicas. Replicas at high temperatures were mainly unfolded, replicas at low temperatures were mostly in the native state, but native-unfolded state transitions within a single replica occurred rarely. Better results were obtained by excluding a randomly chosen subset of water molecules from the solvent interaction scaling,[27] partially negating the benefits of REST compared to the original T-REM.

With REST, the high-temperature replicas are mostly unfolded and thus sample a region in conformational space that does not significantly contribute to the equilibration of the native state. Because of the high entropy of the unfolded state, the probability of refolding to the native state is very low. As shown in Figure 4, partial unfolding also hinders open−closed interconversion rate of T4L. At room temperature, however, the native state is much more stable than the unfolded state. Thus, in REST the only meaningful way for an unfolded protein at high temperature to contribute to the native state at room temperature is by refolding into the native state while traveling down the temperature ladder, which is a very slow process.

The REFT method is based on (i) the notion that only about 0.00001% to 0.01% of polypeptide chains are unfolded at physiological temperatures,[28] and (ii) the assumption that the equilibration rate of proteins is usually limited by slow interdomain dynamics. The REFT potential function tends to keep the rigid domains intact over the full temperature range, while stimulating domain−domain movements at higher temperatures. As such, more simulation time is spent in the native state, which, as is demonstrated in this article, allows the native state to reach an equilibrated ensemble much faster.

It is important to note that the REFT method in its current form cannot be applied a priori to an amino acid sequence of unknown fold. The method is however not meant to study protein folding, but to achieve faster equilibration of the protein native state. Note that protein unfolding is not prevented at any temperature. At higher temperatures however, the fraction of time spent in the unfolded state will generally be small, just as is the case at low temperature with conventional MD.

The choice of the rigid domains will affect the efficiency of the REFT method. If all atoms are considered rigid, the system is equivalent to multiple conventional MD simulations at 300 K. The REST method corresponds to a REFT system with only flexible protein atoms. In this work, we have used the FIRST algorithm to determine the rigid and flexible domains, which derives geometrical constraints from a static protein structure to identify the rigid domains. In the case

of T4L, while large interdomain motions occur, the domains themselves remain intact. This was verified by FIRST calculations on both open and closed X-ray structures, which gave very similar results. For more flexible proteins, a FIRST analysis of multiple protein structures, experimental or from simulation, could be used to distinguish the truly rigid domains from the more flexible ones, as was proposed in the original FIRST article.[16] Other methods could be used to obtain rigidity information, including simple secondary structure identification algorithms, tCONCOORD,[26] which is similarly based on distance constraints, normal mode analysis, and elastic network models.[29]

Independent evidence for the equilibrated state is provided by a study of Goto et al. who obtained approximately equal contributions from open and closed conformers by fitting experimental dipolar couplings to a two-state open−closed model.[30] The superclosed state in part C of Figure 8 is similar to the solution conformation of the T21C/T142C mutant, which contains a disulfide bond between the N- and C-terminal domains.[30] The cleft of this mutant is even more closed than the most closed X-ray T4L conformation (152L, blue triangle in part C of Figure 8), a mutant with three disulfide bonds. Earlier tCONCOORD and multiple conventional MD simulations started at different conformational states did not reveal this compacted state.[26]

The equilibrated average position on the 2D PCA plot (violet diamond, part C of Figure 8) agrees reasonably well with the position of the M6I mutant X-ray structure (150L(c), green square, part C of Figure 8), which resembles the average solution conformation of the T4L homologue C54T/C97A calculated from $^{1}HN-^{15}N$ dipolar couplings.[30] Possible explanations for the difference include: unsatisfactory equilibration, inaccurate force field parameters (the AMBER ff03 force field has been shown to favor α-helices, which might lead to an overly rigid protein,[31] differences between the 150L(c) X-ray structure and the solution structure which were not resolved by the NMR data, and the presence of the two mutations (C54T/C97A) in the NMR structure.

Further improvements of the method are possible. To optimize the balance between rigidity and flexibility, one could refine the scaled potential in eq 4 by explicitly enumerating the atom pairs whose interactions need to be scaled. More accurate rigidity information could be extracted from pregenerated conformational ensembles. Alternatively, using a fast algorithm such as FIRST, rigid domains can be calculated dynamically during the simulation, which would facilitate the formation and destruction of transient domains. Moreover, kinetic information could be extracted from the REM simulations[32] as a way to validate the assumption that the large-scale motion between rigid domains is the slow rate-limiting step.

**Acknowledgment.** We thank B. de Groot for kindly providing the file with reference X-ray structures. S.L.C.M. is a research mandate holder of the Flemish agency for Innovation by Science and Technology (IWT). S.M. is the recipient of a doctoral grant from the Flemish Science Foundation (FWO). This research was conducted utilizing high-performance computational resources provided by the University of Leuven (http://ludit.kuleuven.be/hpc).

## References

(1) Swendsen, R. H.; Wang, J. S. Replica Monte-Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57* (21), 2607–2609.

(2) Hukushima, K.; Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **1996**, *65* (6), 1604–1608.

(3) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141–151.

(4) Frenkel, D.; Smit, B. *Understanding Molecular Simulation. From Algorithms to Applications*; Academic Press: London, U.K., 2002.

(5) Zhou, R. Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (23), 13280–13285.

(6) Garcia, A. E.; Onuchic, J. N. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 13898–13903.

(7) Moors, S. L. C.; Jonckheer, A.; De Maeyer, M.; Engelborghs, Y.; Ceulemans, A. Tryptophan conformations associated with partial unfolding in ribonuclease T1. *Biophys. J.* **2009**, *97* (6), 1778–1786.

(8) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.

(9) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7* (23), 3910–3916.

(10) Moors, S. L. C.; Michielssens, S.; Flors, C.; Dedecker, P.; Hofkens, J.; Ceulemans, A. How is cis-trans isomerization controlled in Dronpa mutants? A replica exchange molecular dynamics study. *J. Chem. Theory Comput.* **2008**, *4* (6), 1012–1020.

(11) Liu, P.; Huang, X. H.; Zhou, R. H.; Berne, B. J. Hydrophobic aided replica exchange: An efficient algorithm for protein folding in explicit solvent. *J. Phys. Chem. B* **2006**, *110* (38), 19018–19022.

(12) Hritz, J.; Oostenbrink, C. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.* **2008**, *128* (14), 144121.

(13) Kannan, S.; Zacharias, M. Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential. *Proteins* **2007**, *66* (3), 697–706.

(14) Zacharias, M. Combining elastic network analysis and molecular dynamics simulations by hamiltonian replica exchange. *J. Chem. Theory Comput.* **2008**, *4* (3), 477–487.

(15) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13749–13754.

(16) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **2001**, *44* (2), 150–165.

(17) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999–2012.

(18) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(19) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(20) Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122.

(21) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.

(22) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pK(a)s and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W368–W371.

(23) de Groot, B. L.; Hayward, S.; van Aalten, D. M.; Amadei, A.; Berendsen, H. J. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins* **1998**, *31* (2), 116–127.

(24) Zhang, X. J.; Wozniak, J. A.; Matthews, B. W. Protein flexibility and adaptability seen in 25 crystalforms of T4 lysozyme. *J. Mol. Biol.* **1995**, *250* (4), 527–552.

(25) Mchaourab, H. S.; Oh, K. J.; Fang, C. J.; Hubbell, W. L. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. *Biochemistry* **1997**, *36* (2), 307–316.

(26) Seeliger, D.; Haas, J.; de Groot, B. L. Geometry-based sampling of conformational transitions in proteins. *Structure* **2007**, *15* (11), 1482–1492.

(27) Huang, X. H.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R. H.; Berne, B. J. Replica exchange with solute tempering: Efficiency in large scale systems. *J. Phys. Chem. B* **2007**, *111* (19), 5405–5410.

(28) Murphy, R. M.; Tsai, A. M. *Misbehaving Proteins. Protein (Mis)folding, Aggregation and Stability*; Springer Science+ Business Media: New York, N.Y., 2006.

(29) Cui, Q.; Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; Chapman & Hall/CRC: Boca Raton, FL, 2006.

(30) Goto, N. K.; Skrynnikov, N. R.; Dahlquist, F. W.; Kay, L. E. What is the average conformation of bacteriophage T4 lysozyme in solution? A domain orientation study using dipolar couplings measured by solution NMR. *J. Mol. Biol.* **2001**, *308* (4), 745–764.

(31) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712–725.

(32) Yang, S.; Onuchic, J. N.; Garcia, A. E.; Levine, H. Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.* **2007**, *372* (3), 756–763.

CT100493V

JCTC Journal of Chemical Theory and Computation

# Long-Range Electrostatic Effects in QM/MM Studies of Enzymatic Reactions: Application of the Solvated Macromolecule Boundary Potential

Tobias Benighaus and Walter Thiel*

*Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany*

Received September 23, 2010

**Abstract:** Long-range electrostatic interactions are important in simulations of enzymatic reactions. They can be divided into the effects due to bulk solvent and those due to the electrostatic potential of the outer macromolecule. We study and quantify the importance of these two effects for two test systems by application of the solvated macromolecule boundary potential (SMBP) [*J. Chem. Theory Comput.* **2009**, *5*, 3114−3128]. We validate the accuracy of the SMBP for these test systems and present a transferable protocol for determination of optimal SMBP parameters as well as recommended default values for these parameters. Two enzymatic reactions with different characteristics are studied: the intramolecular Claisen rearrangement in chorismate mutase that is associated with little charge transfer and the hydroxylation reaction in *p*-hydroxybenzoate hydroxylase that corresponds to a formal "$OH^+$" transfer and thus involves significant charge transfer. It is found that the effects of the electrostatic potential of the outer macromolecule and of bulk solvent are only important in the latter case, where their neglect causes deviations in the computed barriers on the order of 1−2 kcal/mol, respectively. Even larger deviations on the order of several kilocalories per mole are observed for the reaction energies in *p*-hydroxybenzoate hydroxylase if the electrostatic potential of the outer macromolecule is neglected.

## 1. Introduction

Experience in classical simulations of biomolecular systems indicates that a reliable description of long-range electrostatic interactions is of crucial importance.[1−6] Without an accurate description of these effects, simulations can yield results that are even qualitatively wrong. Examples range from the stability of an α helix[7,8] over the thermodynamics of peptide folding processes[9] to properties of lipid bilayers.[10,11]

Efficient approaches to describing electrostatic interactions accurately have been developed for classical force field simulation methods. If all solvent molecules are modeled explicitly, the Ewald summation method in combination with periodic boundary conditions (PBC) is the most established choice.[12−14] Unfortunately, Ewald summation suffers from high computational costs since the enzyme has to be solvated in a water box of adequate size[15−17] to avoid artifacts from the artificially imposed periodicity.[1,18] Alternative approaches include fast multipole methods[19−22] and the use of stochastic boundary conditions.[23−26]

The treatment of long-range electrostatic interactions is far less developed in the context of hybrid quantum mechanical/molecular mechanical (QM/MM) approaches, which are now routinely applied to study enzymatic reactions that necessitate a QM description of bond breaking and forming processes.[6,27,28] The technical difficulties introduced by the QM atoms are the main reason for this situation. Recently, the Ewald summation method has been adapted for QM/MM approaches using semiempirical[29−31] and density functional theory (DFT)[32−34] QM methods.

The region in the enzyme that is described by an accurate QM method in hybrid QM/MM approaches is necessarily small and encompasses at most a few hundred atoms. Therefore, QM/MM methods are best suited for treating

* To whom correspondence should be addressed. E-mail: thiel@mpi-muelheim.mpg.de.

10.1021/ct1005455 
Published on Web 12/03/2010

localized electronic processes. In these cases, boundary potentials are an attractive approach to handling long-range electrostatic interactions.[23−26,35−42] In the boundary potential ansatz, the full system is subdivided into an inner region, which comprises the active site and its neighboring residues, and an outer region, which contains the rest of the enzyme and the outer solvent molecules. While all atoms in the inner region are modeled atomistically, the effect of the outer region atoms on the inner region is mimicked by the boundary potential. Assuming an ideal boundary potential, statistical properties derived from simulations of the inner region in interaction with the boundary potential are the same as those from simulations of the full system. Although an exact and rigorous boundary potential can be constructed by integrating over all outer region degrees of freedom,[42] approximations are indispensable to constructing an efficient implementation.

The generalized solvent boundary potential (GSBP)[43] has been successful in classical force field simulations.[43−48] It was recently adapted for QM/MM approaches using semiempirical QM methods.[49,50] In the GSBP, the outer region solvent molecules are represented by a polarizable dielectric continuum (PDC) and the outer macromolecule region by fixed point charges. The electrostatic potential of the outer region is separated into a static potential, which is induced by the outer region charges being shielded by the PDC, and a dynamic reaction field potential, which is induced by the interaction of the inner region charge distribution with the PDC. The great advantage of the GSBP is its closed-form expression for the dielectric response, which is also valid for irregularly shaped dielectric boundaries. However, construction of this expression is connected with a significant computational overhead.

Therefore, we recently introduced the solvated macromolecule boundary potential (SMBP),[51] which is based on approximations similar to those in the GSBP but has been designed to be efficient in geometry optimizations. Moreover, the SMBP can be used in combination with every QM/MM Hamiltonian.

The effect of long-range electrostatics can be divided into two contributions: interactions with bulk solvent and interactions with the electrostatic potential of the outer macromolecule region (EPOM). By construction, the SMBP allows us to study both effects independently by an appropriate choice of the dielectric constant of the PDC. Here, we apply the SMBP to investigate and quantify the significance of these two effects in QM/MM studies of enzymatic reactions. Two enzymatic reactions were selected as test cases that have been the focus of much theoretical and experimental research: the Claisen rearrangement of chorismate to prephenate in chorismate mutase (CM) and the hydroxylation reaction in *p*-hydroxybenzoate hydroxylase (PHBH). Before applying the SMBP, its accuracy for these two systems is evaluated, optimal values for its inherent parameters are determined, and a transferable protocol to validate the SMBP for other enzymatic systems is presented.

## 2. Methods

In this section, we briefly review the theoretical foundation of the SMBP and the GSBP.

**2.1. Generalized Solvent Boundary Potential.** The GSBP introduces two main approximations with the objective of providing an accurate and efficient approximation of the electrostatic potential, which is induced by the outer macromolecule and solvent region. The outer solvent molecules are replaced by a PDC, and the outer macromolecule region is represented by fixed point charges. Then, the GSBP consists of two terms: the interaction of the inner region charges ($q_A$) with the electrostatic potential of the outer macromolecule being shielded by the PDC ($\phi_s^o$) and the interaction with the self-induced reaction field $\phi_{rf}^i$.

$$\Delta W_{elec}^{GSBP} = \sum_{A\in \text{inner}} q_A \phi_s^o(\mathbf{r}_A) + \sum_{A\in \text{inner}} q_A \phi_{rf}^i(\mathbf{r}_A) \qquad (1)$$

With the outer macromolecule region being fixed, $\phi_s^o$ is constant and needs to be computed only once at the beginning of a molecular dynamics (MD) simulation. Direct computation of the second term, however, is prohibitively expensive in MD runs since the reaction field potential has to be updated for each configuration, i.e., in each step.

To circumvent repeated solution of the Poisson−Boltzmann (PB) equation to update $\phi_{rf}^i$, a Green's function has been introduced to express the inner reaction field potential for a given charge density of the inner region ($\rho_i$).[43]

$$\phi_{rf}^i(\mathbf{r}) = \int d\mathbf{r}'\, \rho_i(\mathbf{r}')\, G_{rf}(\mathbf{r}, \mathbf{r}') \qquad (2)$$

The reaction field Green's function and the inner region charge density are projected onto the same basis set $\{b_n\}$. The solvation energy of the inner region can then be calculated from the reaction field matrix, $M_{rf}$, and the generalized multipole moments of the inner region charge density ($Q_n$).

$$\Delta W_{elec}^{GSBP} = \sum_{A\in \text{inner}} q_A \phi_s^o(\mathbf{r}_A) + \frac{1}{2}\sum_{mn} Q_m M_{mn} Q_n \qquad (3)$$

In this way, the GSBP offers an analytical expression for the electrostatic potential and avoids solution of the PB equation in each MD step. However, it is important to point out that application of the GSBP is connected with a significant overhead since computation of the reaction field matrix requires solving the PB equation a few hundred times.[43,50]

**2.2. Solvated Macromolecule Boundary Potential.** Although free energy computations are becoming more popular in the QM/MM field,[28,52−55] the majority of QM/MM studies still rely on (constrained) geometry optimizations to compute potential energy differences and molecular structures of stationary points. Hence, we developed the solvated macromolecule boundary potential (SMBP) that targets QM/MM geometry optimizations with ab initio or DFT QM methods. Moreover, conceptual proximity to the GSBP was a requirement to allow application of the GSBP in free energy perturbation studies on the basis of QM/MM/SMBP potential energy profiles.[51]

The SMBP relies on the same basic approximations as the GSBP: The outer macromolecule region is represented by fixed point charges and the outer solvent molecules by the PDC. However, the SMBP differs in the computation of the inner self-induced reaction field potential, which is updated in each geometry optimization step by solving the PB equation. This is more efficient than the initial construction of the reaction field matrix by repeated solution of the PB equation (typically about 800 times).[50]

In addition to efficiency in geometry optimizations, the SMBP was designed to be applicable with all QM/MM Hamiltonians. To achieve this, the interaction with the static outer region potential and the reaction field potential is expressed as the interaction of the QM and MM charge densities with the effective potential that they experience, respectively.

$$\Delta W_{\mathrm{elec}}^{\mathrm{SMBP}} = \int \mathrm{d}\mathbf{r}\, \rho_{\mathrm{QM}}(\mathbf{r})\, \phi_{\mathrm{tot}}^{\mathrm{QM}}(\mathbf{r}) + \int \mathrm{d}\mathbf{r}\, \rho_{\mathrm{MM}}\, \phi_{\mathrm{tot}}^{\mathrm{MM}}(\mathbf{r}) \tag{4}$$

with

$$\phi_{\mathrm{tot}}^{\mathrm{QM}}(\mathbf{r}) = \phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}) + \phi_{\mathrm{rf}}^{\mathrm{MM}}(\mathbf{r}) + \frac{1}{2}\phi_{\mathrm{rf}}^{\mathrm{QM}}(\mathbf{r}) \tag{5}$$

$$\phi_{\mathrm{tot}}^{\mathrm{MM}}(\mathbf{r}) = \phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}) + \frac{1}{2}\phi_{\mathrm{rf}}^{\mathrm{MM}}(\mathbf{r}) \tag{6}$$

Since the QM and MM reaction field potentials, $\phi_{\mathrm{rf}}^{\mathrm{QM}}(\mathbf{r})$ and $\phi_{\mathrm{rf}}^{\mathrm{MM}}(\mathbf{r})$, depend on the configuration of the inner region, they have to be updated in each step. Computation of the QM reaction field potential is further exacerbated by the mutual dependence of the QM charge density and the QM reaction field potential. To solve this problem, a self-consistent reaction field (SCRF) procedure is applied: the wave function and the effective QM potential are updated alternately, with the QM charge density being fixed during computation of $\phi_{\mathrm{tot}}^{\mathrm{QM}}$ and vice versa. This approach enables a flexible interface of SMBP and QM codes: ESP charges are used to represent the QM density in the PB equation, and a small set of virtual surface charges is used to represent the SMBP during optimization of the wave function. Hence, the SMBP can be used in combination with virtually any QM code that is suitable for QM/MM calculations.[51]

The QM/MM free energy perturbation (QM/MM-FEP) approach introduced by Yang et al.[56] allows the computation of QM/MM free energy profiles using ab initio or DFT QM methods at moderate computational costs. QM/MM-FEP is based on three assumptions: (1) Sampling over the QM and MM degrees of freedom can be separated in a product ansatz, (2) finite-temperature effects in the QM region can be estimated using the harmonic approximation, and (3) the electrostatic interactions of the MM region with the QM atoms can be approximated as interactions of MM point charges with fixed ESP point charges on the QM atoms. The QM/MM-FEP calculation provides a free energy correction to the potential energy profile along a predefined reaction coordinate.

The SMBP employs the same approximations as the GSBP. Therefore, the SMBP and GSBP can be used complementarily. If one uses a QM/MM/SMBP Hamiltonian to compute the potential energy profile, the efficient GSBP can be applied to sample over the MM degrees of freedom in the subsequent FEP step. The resulting QM/MM/GSBP-FEP method reduces the computational costs of the FEP step by typically 1 order of magnitude in the cases studied.[51]

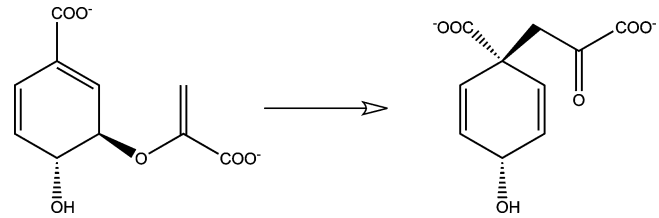



***Figure 1.*** Intramolecular Claisen rearrangement catalyzed by chorismate mutase.

## 3. Computational Details

GSBP and SMBP have been implemented in a developmental version of the modular program package ChemShell.[57,58] In the present study, the QM energies and gradients were evaluated with the MNDO[59] and Turbomole 5.7.1 programs.[60] The DL_POLY[61] code was employed to run the CHARMM22 force field in all calculations of the MM part.[62] Hydrogen link atoms were applied in combination with the charge-shift scheme[57] to saturate the QM region. The HDLCOpt optimizer was used to optimize stationary points in hybrid delocalized internal coordinates.[63] The PB equation was solved with the ChemShell PB module, which applies the optimal successive over-relaxation method in combination with Gauss-Seidel relaxation to compute the electrostatic potential.[64,65] A maximum absolute change in every grid point of $2 \times 10^{-5}$ au was adopted as the convergence criterion, and third-order B splines were used to interpolate between the grid points.[66] The dielectric boundary was defined as the superposition of the van der Waals radii from the CHARMM22 force field. MD simulations were performed unter NVT conditions with a time step of 1 fs at a temperature of 300 K controlled by a Nosé−Hoover chain thermostat.[67−70] Free water molecules were kept rigid with SHAKE constraints,[71] and the mass of deuterium was assigned to all hydrogen atoms. The QM reaction field potential was considered converged when the root-mean-squared deviation was below $2 \times 10^{-5}$ au. At the beginning of the SCRF procedure, a neutral charge was assigned to all QM atoms.

## 4. Results

**4.1. Chorismate Mutase: Validation and Parameter Determination.** CM catalyzes the intramolecular Claisen rearrangement from chorismate to prephenate (Figure 1). A recent review provides a summary of theoretical work on the reaction mechanism and the origin of catalysis in CM.[28] The Claisen rearrangement is a pericyclic reaction without significant charge transfer, and one would thus expect only minor effects of the EPOM and bulk solvent on this enzymatic reaction.

The setup for CM was based on a system that has been used in previous work.[52] Initial coordinates were taken from

the crystallographic structure of *Bacillus subtilis* CM (PDB code 2CHT) with a bound transition state analog (TSA). Only the first of four trimers in the asymmetric unit was retained, and the TSA between chains A and B was transformed into a chorismate molecule. The other TSAs were removed. The system was solvated with a 30 Å water sphere and then subjected to a 200 ps QM/MM MD simulation using self-consistent charge-density functional tight binding (SCC-DFTB)[72] as a QM method. Our setup started from the resulting structure corresponding to snapshot 1 in the previous study.[52] To generate 10 initial configurations for the QM/MM calculations that are independent of the previous study, snapshots were taken every 10 ps after 100 ps of extra equilibration. The QM region consisted only of the chorismate molecule, which was modeled by the semiempirical AM1 Hamiltonian.[73] All protein atoms were assigned to the MM region. For each configuration, the inner region was centered on the initial position of the $C_1$ atom (following IUPAC nomenclature) with an extended dielectric cavity radius of 21.0 Å. All atoms within 19.0 Å of the center were assigned to the inner region and modeled explicitly. Due to the inaccuracies of the SMBP and the GSBP at the boundary,[50,51] the inner region was further subdivided into an active inner region and a frozen inner region. All atoms within 17.0 Å of the center belong to the active region so that it is surrounded by an "insulation" region of 2 Å. A spherical restraint with a radius of 17.0 Å and a force constant of 50 kcal/(mol Å$^2$) was applied to the oxygen atoms of all active water molecules in MD calculations with the QM/MM/GSBP method. The reaction is described by means of a reaction coordinate (RC) defined as the difference of the lengths of the breaking C−O and the forming C−C bonds. Potential energy profiles of the reaction are computed by constraining the RC to values from −2.5 to +2.5 Å in steps of 0.1 Å.

The electrostatic potentials that constitute the SMBP (see eqs 5 and 6) are obtained as grid-based solutions of the PB equation. A focusing approach is applied to allow usage of fine grids for the inner region.[74] The PB equation is first solved for a coarse outer grid that covers the full system and then for a fine inner grid that focuses on the inner region. The boundary values of the inner grid are set by interpolation from the outer grid. Therefore, the mesh sizes of the grids are the main parameters that determine the accuracy as well as the efficiency of the SMBP. They have to be chosen carefully. Hence, one of the objectives of this study is the development of a transferable protocol to estimate adequate mesh sizes based only on fast single-point energy and gradient calculations.

In a vacuum environment, that is, with a dielectric constant of 1 anywhere in space, the electrostatic potential of the SMBP and of exact Coulombic electrostatics has to be identical. This allows simple evaluation of the accuracy of the SMBP by direct comparison to standard QM/MM results. Table 1 provides the mean absolute (MAD) and maximum absolute deviations (MAX) of the gradient components for mesh size combinations of 0.15, 0.25, 0.4, 0.6, and 0.8 Å for the inner grid and 0.8, 1.25, 1.5, 2.0, and 2.5 Å for the outer grid. Here, all atoms within 18 Å of the center were considered to account for fluctuations of the active atoms that may occur later in the simulations. MAD and MAX results show no significant dependence on the mesh size of the outer grid but a strong dependence on the mesh size of the inner grid. For all mesh size combinations, the MAD values are between $10^{-5}$ and $5 \times 10^{-5}$ au. The default convergence criterion for QM/MM geometry optimizations is a maximum absolute gradient component of less than $4.5 \times 10^{-4}$ au.[63] From this perspective, the MAD results can be deemed accurate. However, they do not allow determination of the optimal mesh size. The MAX deviations are more helpful for this task. For inner mesh sizes of ≤0.25 Å, MAX deviations are below the convergence criterion so that this mesh size excels as a safe choice at acceptable computational costs. Since neither the accuracy nor the computational demands depend strongly on the mesh size of the outer grid, a relatively fine outer grid with a mesh size of 1.25 Å was selected in combination with an inner grid spacing of 0.25 Å for all calculations (unless noted otherwise).

***Table 1.*** Mean Absolute (MAD) and Maximum Absolute (MAX) Deviations [$10^{-4}$ au] of the Electrostatic Forces Computed with the SMBP for the Chorismate Mutase System and Averaged over 10 Configurations (Relative to QM/MM Results with Full Coulombic Electrostatics)[a]

| | inner grid size [Å] | | | | |
|---|---|---|---|---|---|
| outer grid size [Å] | 0.15 | 0.25 | 0.40 | 0.60 | 0.80 |
| | MAD—atoms within 18 Å | | | | |
| 0.80 | 0.11 | 0.12 | 0.19 | 0.31 | 0.45 |
| 1.25 | 0.16 | 0.16 | 0.23 | 0.34 | 0.48 |
| 1.50 | 0.17 | 0.17 | 0.23 | 0.35 | 0.48 |
| 2.00 | 0.15 | 0.16 | 0.22 | 0.34 | 0.47 |
| 2.50 | 0.18 | 0.18 | 0.24 | 0.36 | 0.49 |
| | MAX—atoms within 18 Å | | | | |
| 0.80 | 2.19 | 3.67 | 8.46 | 16.29 | 20.99 |
| 1.25 | 3.15 | 4.01 | 8.58 | 16.40 | 21.12 |
| 1.50 | 3.37 | 4.04 | 8.59 | 16.44 | 21.07 |
| 2.00 | 2.94 | 4.01 | 8.50 | 16.37 | 21.09 |
| 2.50 | 3.26 | 4.22 | 8.58 | 16.37 | 21.17 |

[a] Different mesh size combinations were used.

Detailed examination shows that the accuracy of the SMBP strongly depends on the radial position of the atoms. This point is illustrated in Figure 2, which shows the MAD and MAX values for different inner grid mesh sizes as a function of the radial position. In the center of the inner region, the electrostatic potential varies only slowly and is described accurately by all tested mesh sizes. At the boundary, however, the electrostatic potential becomes more complex and the deviation increases significantly for mesh sizes >0.4 Å. Therefore, fine inner mesh sizes appear to be necessary. Since the chemical process occurs in the center of the inner region, however, it is possible that moderate deviations at the boundary are tolerable. To test this hypothesis, reaction energies and activation energies were computed for grid size combinations ranging from 0.25/1.25 Å to 1.8/3.5 Å. The results are documented in Table S1 of the Supporting Information (SI). In comparison to full Coulombic electrostatics, the MAD and MAX deviations of the potential energy differences are below 0.3 and 0.8 kcal/mol, respectively, if the inner mesh size is ≤0.8 Å. Such an inner grid mesh size

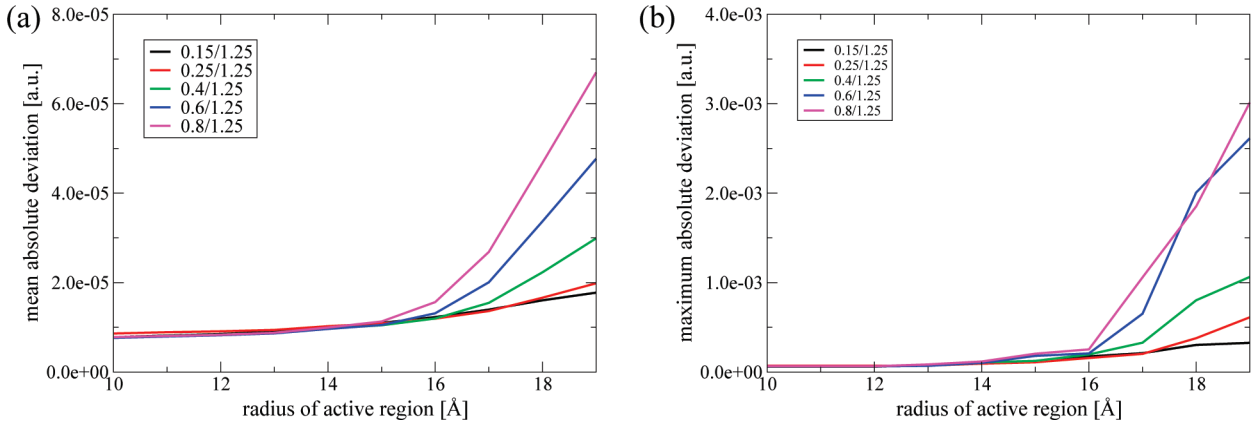



***Figure 2.*** Mean absolute deviations (a) and maximum absolute deviations (b) of the electrostatic forces at all atoms inside the active region relative to the exact QM/MM values for the chorismate mutase (CM) test system. Results are shown for different mesh sizes of the inner grid and are plotted as a function of the radius of the active region. An outer grid size of 1.25 Å is used. All calculations were performed on configuration 1 of the CM system. The radius of the inner region was 19 Å (see text).

corresponds to a MAX deviation of the gradient components in the range of $2 \times 10^{-3}$ au (see Table 1). We conclude that the SMBP will provide results of high accuracy if the mesh size is chosen such that the MAX values do not exceed $4.5 \times 10^{-4}$ au anywhere in the inner region. The SMBP will still give results of adequate accuracy with coarser mesh sizes if the MAX values do not exceed $4.5 \times 10^{-4}$ au for atoms more than 3 Å away from the boundary and if the MAX values do not exceed $2 \times 10^{-3}$ au for atoms more than 1 Å away from the boundary.

In the GSBP, another important parameter has to be determined: the size of the basis set which models the inner region charge distribution. The chosen orthonormal basis functions are based on spherical harmonics so that the size of the basis set is determined by the order of the highest multipole moment that is included. Table 2 gives the MAD and MAX deviations for basis sets of increasing size determined by maximum multipole moments from order $L = 1$ to $L = 19$. Moreover, it provides the fraction of inaccurate gradient components whose deviation is larger than the standard convergence criterion (see above). SMBP values serve as the reference, since the SMBP represents the basis set limit of the GSBP. In previous applications of the GSBP, multipole moments up to order $L = 19$ were usually included. The MAX results confirm this choice and show that with $L = 19$, MAX deviations on the order of $4.5 \times 10^{-4}$ au are observed, which is sufficiently accurate. The fractions of inaccurate gradient components indicate that the reaction field potential converges in this system at $L = 17$. The residual error can be attributed to technical differences of SMBP and GSBP. It seems interesting to check whether looser criteria can also yield results of adequate accuracy. Table S2 in the SI shows that the basis-set-dependent error converges for $L = 10$ if only those atoms are considered that are at least 3 Å away from the boundary. The impression that $L = 10$ provides sufficient accuracy is further supported by the fact that the MAX deviation for all atoms at least 1 Å away from the boundary is $1.3 \times 10^{-3}$ au (Table 2) and, therefore, below the criterion of $2 \times 10^{-3}$ au that was used to determine the mesh sizes. Hence, $L = 19$ emerges as the safe and $L = 10$ as the efficient choice. In section 4.2, both values are tested, and it will be shown that they describe bulk solvent effects in CM with similar accuracy.

***Table 2.*** Mean Absolute (MAD) and Maximum Absolute (MAX) Deviations [$10^{-4}$ au] and the Fraction of Inaccurate Gradient Components ($x_{grad}$) [%] of the Electrostatic Forces of the Chorismate Mutase System Computed with the GSBP with Different Basis Set Sizes (Relative to SMBP Results, See Text)[c]

| L[a] | MAD | MAX | $x_{grad}$[b] |
|---|---|---|---|
| 1 | 4.00 | 69.77 | 26.47 |
| 2 | 3.19 | 60.99 | 21.03 |
| 3 | 2.39 | 50.28 | 15.60 |
| 4 | 1.74 | 39.47 | 11.05 |
| 5 | 1.22 | 30.00 | 6.76 |
| 6 | 0.95 | 27.21 | 4.67 |
| 7 | 0.76 | 22.59 | 3.16 |
| 8 | 0.61 | 18.56 | 1.97 |
| 9 | 0.50 | 14.48 | 1.13 |
| 10 | 0.42 | 13.31 | 0.70 |
| 11 | 0.36 | 10.97 | 0.40 |
| 12 | 0.32 | 8.75 | 0.25 |
| 13 | 0.29 | 7.34 | 0.12 |
| 14 | 0.27 | 5.95 | 0.07 |
| 15 | 0.25 | 5.26 | 0.05 |
| 16 | 0.24 | 5.03 | 0.04 |
| 17 | 0.23 | 4.72 | 0.03 |
| 18 | 0.22 | 4.55 | 0.03 |
| 19 | 0.22 | 4.53 | 0.03 |

[a] highest order multipole moment. [b] fraction of gradient components with a deviation >4.5 × $10^{-4}$ a.u. [%]. [c] All values are averaged over 10 configurations and computed with a mesh size combination of 0.25/1.25 Å.

**4.2. Chorismate Mutase: Results.** We now apply the SMBP to study the effects of the EPOM and bulk solvent on the enzymatic reaction in CM. Table 3 gives the reaction energies and activation energies that were computed with different methods to describe these effects. In each case, the energy differences were averaged over 10 configurations. The standard deviation of any given mean value defines a confidence interval corresponding to a confidence limit of 68%. Statistically significant are differences between mean values that are larger than the confidence interval of the reference value.

Standard QM/MM calculations with full Coulombic electrostatics (Coulomb in Table 3) give reaction and

***Table 3.*** Reaction Energies ($\Delta E$) and Activation Energies ($\Delta E^{\ddagger}$) of the Claisen Rearrangement in Chorismate Mutase [kcal/mol] Computed with Different Treatments of Long-Range Electrostatics

| | Coulomb | | SMBP(vac)[a] | | NOR[b] | | SMBP(solv)[c] | | SMBP(solv,app)[d] | |
|---|---|---|---|---|---|---|---|---|---|---|
| configuration | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ |
| 1 | −20.0 | 33.0 | −19.3 | 33.1 | −19.9 | 33.4 | −20.1 | 32.6 | −19.1 | 32.8 |
| 2 | −17.3 | 34.8 | −17.2 | 34.8 | −11.5 | 41.6 | −18.3 | 33.1 | −17.2 | 34.3 |
| 3 | −17.5 | 40.1 | −17.1 | 40.4 | −18.9 | 37.2 | −16.8 | 39.5 | −16.5 | 40.2 |
| 4 | −17.7 | 38.2 | −17.7 | 38.3 | −14.9 | 36.5 | −17.3 | 37.5 | −17.3 | 37.8 |
| 5 | −21.9 | 32.4 | −21.8 | 32.5 | −21.8 | 33.1 | −21.9 | 32.4 | −21.9 | 32.7 |
| 6 | −18.0 | 36.2 | −17.8 | 36.3 | −17.1 | 36.6 | −17.7 | 35.4 | −17.7 | 35.4 |
| 7 | −19.6 | 35.3 | −19.5 | 35.4 | −19.3 | 35.3 | −19.9 | 34.7 | −19.6 | 35.1 |
| 8 | −19.3 | 35.8 | −19.3 | 35.7 | −19.2 | 37.2 | −19.4 | 34.9 | −19.0 | 36.0 |
| 9 | −16.3 | 34.4 | −16.1 | 34.5 | −17.4 | 34.7 | −17.9 | 33.5 | −17.2 | 34.0 |
| 10 | −20.7 | 30.6 | −20.7 | 30.6 | −20.6 | 31.3 | −21.0 | 30.4 | −20.0 | 31.5 |
| mean value | −18.8 | 35.1 | −18.6 | 35.2 | −18.1 | 35.7 | −19.0 | 34.4 | −18.5 | 35.0 |
| standard deviation of data[e] | 1.8 | 2.8 | 1.8 | 2.8 | 3.0 | 2.9 | 1.7 | 2.6 | 1.7 | 2.6 |
| standard deviation of mean[f] | 0.6 | 0.9 | 0.6 | 0.9 | 0.9 | 0.9 | 0.5 | 0.8 | 0.5 | 0.8 |
| MAD[g] | | | 0.2 | 0.1 | 1.3 | 1.5 | | | 0.5 | 0.6 |

[a] $\varepsilon = 1$. [b] Neglect of outer region. [c] $\varepsilon = 80$. [d] $\varepsilon = 80$, $\phi_{rf}^{QM} = 0$. [e] Standard deviation of individual energy values. [f] Standard deviation of the mean value (68% confidence limit). [g] Mean absolute deviation relative to full Coulombic electrostatics. For SMBP(solv,app), SMBP(solv) values are used as a reference.

activation energies of −18.8 and +35.1 kcal/mol, respectively. In the standard QM/MM approach using electronic embedding, the EPOM is computed without approximation, but the effect of the bulk solvent is neglected. Compared to the experimental value of 12.7 kcal/mol,[75] the computed value for the activation energy is strongly overestimated because of the use of the AM1 Hamiltonian in the QM region.[28] Despite this shortcoming, we expect the AM1 Hamiltonian to be adequate for studying the relative effects caused by long-range electrostatics, which should be captured at this level in a realistic manner. For two of the 10 configurations, the conclusions drawn from AM1 results were confirmed by more accurate DFT calculations using the B3LYP functional[76−78] in combination with a 6-31G* basis (see below).

When the SMBP is applied under vacuum conditions (SMBP(vac)), the effect of the EPOM is approximated and bulk solvent effects are neglected. The SMBP describes the EPOM accurately and reproduces the standard QM/MM results with very small deviations. The mean values for the reaction and activation energy differ by only 0.2 and 0.1 kcal/mol, respectively. Also, the results for the individual configurations are very similar, as indicated by MAD values of similar size.

Going one step further, one could completely neglect the electrostatic influence from the outer solvent and macromolecule region (neglect of outer region, NOR). This simplest approximation should lead to significant deviations from the standard QM/MM results, if the energetics are sensitive to long-range electrostatics. With the NOR approximation, the mean reaction and activation energies increase by 0.7 and 0.6 kcal/mol, respectively (Table 3). These differences are at the boundary of the confidence interval of the standard QM/MM results and thus cannot be considered significant. However, the MAD values of 1.3 and 1.5 kcal/mol for reaction and activation energies show that the results deviate significantly for the individual configurations. With less fortuitous error cancellation, significant discrepancies of more than 1 kcal/mol are therefore possible when applying the NOR approximation.

***Table 4.*** Reaction Energies ($\Delta E$) and Activation Energies ($\Delta E^{\ddagger}$) in Chorismate Mutase [kcal/mol] Computed at the QM(B3LYP/6- 31G*)/MM Level of Theory with Different Treatments of Long-Range Electrostatics (See Table 3 for Notation)

| | Coulomb | | SMBP(vac) | | NOR[a] | | SMBP(solv) | |
|---|---|---|---|---|---|---|---|---|
| configuration | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ |
| 3 | −14.0 | 15.9 | −14.8 | 15.5 | −14.3 | 14.8 | −13.5 | 15.7 |
| 8 | −16.2 | 13.0 | −16.2 | 13.0 | −17.1 | 12.8 | −16.3 | 12.6 |

[a] Neglect of outer region.

The effect of bulk solvent on this reaction was studied using the SMBP with a dielectric constant of 80 for the outer solvent region (SMBP(solv)). The inclusion of bulk solvent lowers the reaction energy only marginally by 0.2 kcal/mol. A stronger change is observed for the activation energy, which is reduced by 0.7 kcal/mol to a value of 34.4 kcal/mol. However, both values lie still within the statistical error bars of the standard QM/MM results so that the effect of bulk solvent has to be deemed insignificant for this reaction.

For configurations 3 and 8, the effects of the EPOM and bulk solvent were checked by DFT/MM calculations using B3LYP/6-31G* for the QM region. The results are summarized in Table 4. Focusing on the computed barriers, we first note that the standard DFT/MM treatment with Coulombic electrostatics yields values (13.0−15.9 kcal/mol) close to experimental results (see above). These barriers are lowered slightly when applying the SMBP under vacuum conditions (by 0.0−0.4 kcal/mol) and more so when using the NOR approximation (by 0.2−1.1 kcal/mol). The barrier lowering due to bulk solvent is again small (0.2−0.4 kcal/mol). These DFT results are fully consistent with the AM1/MM results (Table 3).

Since the QM region is usually located far away from the dielectric boundary, one can assume that the QM contribution to the reaction field potential is small and can be neglected. This is the SMBP(solv,app) approximation which neglects the $\phi_{rf}^{QM}$ term in eq 5 and thus avoids the SCRF procedure, thereby accelerating the SMBP calculations. This approach yields reaction and activation energies of −18.5 and +35.0

***Table 5.*** Reaction Free Energies ($\Delta A$) and Activation Free Energies ($\Delta A^{\ddagger}$) in Chorismate Mutase [kcal/mol] at $T = 300$ K Computed with Different Treatments of Long-Range Electrostatics

| | GSBP(vac) | | GSBP(solv)[a] | | GSBP(solv,fast)[b] | | NOR[c] | |
|---|---|---|---|---|---|---|---|---|
| configuration | $\Delta A$ | $\Delta A^{\ddagger}$ | $\Delta A$ | $\Delta A^{\ddagger}$ | $\Delta A$ | $\Delta A^{\ddagger}$ | $\Delta A$ | $\Delta A^{\ddagger}$ |
| 1 | −19.0 | 31.1 | −17.6 | 31.1 | −17.0 | 31.6 | −16.9 | 31.7 |
| 2 | −17.1 | 32.6 | −16.2 | 33.1 | −15.7 | 33.1 | −14.5 | 36.8 |
| 3 | −16.0 | 37.8 | −14.6 | 36.2 | −15.6 | 36.6 | −16.3 | 33.9 |
| 4 | −18.0 | 35.1 | −15.2 | 35.7 | −15.9 | 34.6 | −15.8 | 33.7 |
| 5 | −18.3 | 33.5 | −19.2 | 32.2 | −18.3 | 32.7 | −18.9 | 32.2 |
| 6 | −17.0 | 34.8 | −17.0 | 32.4 | −16.6 | 33.4 | −15.8 | 34.6 |
| 7 | −18.6 | 33.2 | −18.1 | 32.5 | −18.5 | 32.1 | −19.1 | 31.7 |
| 8 | −18.1 | 33.2 | −16.6 | 33.4 | −17.3 | 32.9 | −16.2 | 33.1 |
| 9 | −17.5 | 32.2 | −16.6 | 32.3 | −17.3 | 31.7 | −16.2 | 31.9 |
| 10 | −19.7 | 30.4 | −18.4 | 30.2 | −18.0 | 31.1 | −17.8 | 31.9 |
| mean value | −17.9 | 33.4 | −16.9 | 32.9 | −17.0 | 33.0 | −16.8 | 33.1 |
| standard deviation of data[d] | 1.1 | 2.1 | 1.4 | 1.8 | 1.1 | 1.6 | 1.4 | 1.6 |
| standard deviation of mean[e] | 0.3 | 0.7 | 0.5 | 0.6 | 0.3 | 0.5 | 0.5 | 0.5 |

[a] Grid size: 0.25/1.25 Å, 400 basis functions ($L = 19$). [b] Grid size: 0.6/2.0 Å, 121 basis functions ($L = 10$). [c] Neglect of outer region. [d] Standard deviation of individual energy values. [e] Standard deviation of the mean value (68% confidence limit).

kcal/mol, respectively. These values are within the statistical error bars of the standard QM/MM and the SMBP(solv) results. Hence, this approximation is valid for this reaction.

Finally, we checked whether the effect of the EPOM and bulk solvent is more pronounced for the dynamical behavior of this system, that is, when computing free energy differences. We applied the QM/MM-FEP method to compute the free energies of reaction and activation at $T = 300$ K (Table 5). Along the RC, the reaction was split up into discrete windows. For each of these, ESP charges for the QM atoms were derived by fitting to the electrostatic potential at the 200 MM atoms closest to the QM region. For each window, the active region was equilibrated for 10 ps with the QM region held fixed and the QM-MM electrostatic interactions modeled classically using the ESP charges. Subsequently, energy differences were sampled for 10 ps, and the data were subjected to statistical tests for lack of correlation and trend.[79] If necessary, data were discarded to obtain a series of values without a trend. Although the effective sampling length was determined by the statistical tests, it was ensured that at least 5 ps of data were retained.

In the QM/MM-FEP calculations, the GSBP was applied to sample the MM phase space efficiently.[51] Again, the accuracy of the QM/MM/GSBP-FEP ansatz was validated against standard QM/MM-FEP results using vacuum conditions. However, since QM/MM-FEP calculations are computationally intense, free energy profiles without GSBP were obtained only for two configurations. The results are documented in Table S3 of the SI and show that QM/MM/GSBP reproduces QM/MM results in FEP calculations well, with deviations of 0.0−0.2 kcal/mol in the reaction free energies ($\Delta A$) and 0.3−0.5 kcal/mol in the activation free energies ($\Delta A^{\ddagger}$). We now consider the changes in the calculated mean values when going from potential energies (Table 3) to free energies (Table 5). Using the GSBP in a vacuum, the reaction energy increases by +0.7 to −17.9 kcal/mol while the activation energy decreases by 1.8 to 33.4 kcal/mol. Four terms contribute to the difference between free and potential energy: the zero point vibrational energy (ZPE, $\Delta E_{ZPE}$), the thermal contribution to the internal energy ($\Delta U_{th}$), the QM entropy contribution ($-T\Delta S_{QM}$), and the entropy of the QM−MM interactions ($-T\Delta S_{QM-MM}$). In the QM/MM-FEP approach, we assume that $-T\Delta S_{QM-MM}$ is the difference of free and potential QM−MM interaction energies and neglect other contributions. Table 6 shows that the total effect on the barrier is dominated by the ZPE. The entropic QM contribution from the harmonic approximation (0.8 kcal/mol) cancels the entropic QM−MM contribution from the sampling (−0.8 kcal/mol) so that entropy does not contribute significantly to finite-temperature effects on the activation energy. This result contradicts the experimental observation of an entropic contribution of $(-T\Delta S)_{exp} = 2.7$ kcal/mol.[75] The negative change of entropy in the transition state has been ascribed to the loss of conformational flexibility due to the partial formation of two covalent bonds.[52] Since the degrees of freedom of the QM region are held fixed during QM/MM-FEP sampling, this phenomenon cannot be captured in the free energy difference of QM−MM interactions. The entropic QM contribution based on the harmonic approximation shows the correct trend but underestimates the magnitude. Therefore, one may tentatively assume that a significant contribution to the negative change in entropy in the activation energy of CM comes from the degrees of freedom that involve coupled motions of QM and MM atoms which are not sampled in the QM/MM-FEP ansatz. The deviation of QM/MM/GSBP-FEP results from experimental results would then arise from the QM/MM-FEP ansatz itself, not from the approximations in the boundary potentials. This view is supported by the results of previous semiempirical QM/MM umbrella sampling simulations (QM = SCC-DFTB) that do not restrict the flexibility of the QM region and reproduce the entropic

***Table 6.*** Contributions [kcal/mol] to the Differences between Free and Potential Energies in Chorismate Mutase Using the QM/MM/GSBP-FEP Method (See Text)[a]

| | reaction | activation |
|---|---|---|
| $\Delta E_{ZPE}$ | 0.3 | −1.5 |
| $\Delta U_{th}$ | −0.1 | −0.3 |
| $-T\Delta S_{QM}$ | 0.3 | 0.8 |
| $-T\Delta S_{QM-MM}$ | 0.2 | −0.8 |
| total | 0.7 | −1.8 |

[a] All values are averaged over 10 configurations.

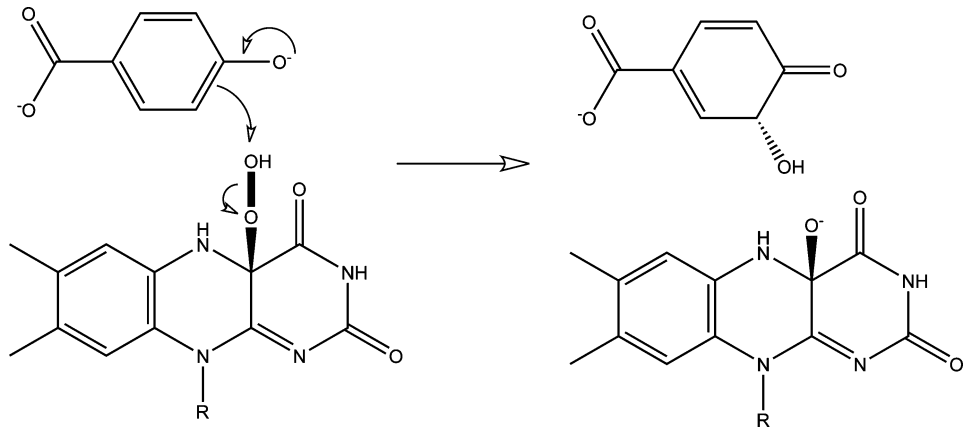


***Figure 3.*** Hydroxylation reaction catalyzed by *p*-hydroxybenzoate hydroxylase. R denotes the ribityl side chain of the hydroperoxo flavine-adenine cofactor.

contribution to the barrier with good accuracy.[52] It is further substantiated by the observation that QM/MM-FEP predicts a negative entropic QM−MM contribution ($-T\Delta S_{QM-MM}$) to the activation free energy independent of the boundary potential and the QM/MM level of theory (Table S4, SI).

The effects of bulk solvent on the free energy differences were studied using a dielectric constant of 80 for the outer solvent region (GSBP(solv)). The activation free energy decreases upon such inclusion of bulk solvent by 0.5 kcal/mol and is thus still within the confidence interval of the vacuum result. The reaction energy, however, increases significantly by +1.0 to −16.9 kcal/mol. These results are reproduced quite accurately if coarser mesh sizes of 0.6/2.0 Å and a smaller basis set ($L = 10$) are used (GSBP-(solv,fast)). However, simply neglecting the electrostatic effect of the outer region (NOR) yields results of very similar accuracy. This can be explained by the shielding effect of the outer solvent, which is mimicked most simply by neglecting the outer region charges. The quantitative accuracy is probably fortuitous given that the results for the potential energy differences deviate significantly from the SMBP(solv) results (Table 3).

On the basis of these results, we may draw the following conclusions for CM: The SMBP and the GSBP reproduce long-range electrostatic interactions with high accuracy when using fine grids with mesh sizes of 0.25/1.25 Å as well as coarser grids with mesh sizes of 0.6/2.0 Å. It should be emphasized, however, that the influence of the EPOM on this reaction is limited. This can be attributed to the "neutral" character of the intramolecular Claisen rearrangement in CM. Neglecting the electrostatic influence of the outer region does not have much effect on the mean potential energy differences although the results for the individual configurations differ significantly. Similarly, bulk solvent effects do not affect the potential energy differences. A statistically significant influence of bulk solvent on the order of 1 kcal/mol was only observed for the reaction free energy.

**4.3. *p*-Hydroxybenzoate Hydroxylase.** The hydroxylation reaction in the catalytic cycle of PHBH has been the focus of many theoretical investigations. It has become a prototypical test system to benchmark theoretical methods for computational enzymology. A summary of theoretical work on PHBH is provided in a recent review.[28] In this reaction, a formal $OH^+$ unit is transferred from the flavin-adenine hydroperoxide cofactor (FADHOOH) to the *p*-hydroxybenzoate substrate (pOHB; see Figure 3). Since the reaction is associated with a considerable charge transfer, it is a potentially good test case for reactions with stronger long-range electrostatic interactions.

The PHBH setup was based on a system that has been used in previous studies.[79−81] It was generated by solvating the enzyme containing the FADHOOH cofactor and the pOHB substrate in a 90 Å solvent box and equilibrating it with gradually decreasing harmonic restraints on the non-water atoms. In a following MD run, harmonic restraints were acting only on the cofactor and substrate, and in the resulting structure, which served as starting point for our setup, all water molecules outside 11 Å from any protein atom were discarded.

Due to a change of force field from GROMOS (previously) to CHARMM (this study), the system was re-equilibrated for 500 ps with constraints on the cofactor, substrate, and all water molecules with oxygen atoms outside 2.9 Å from any protein atom. Five configurations were selected from this MD run after 420, 440, 460, 480, and 500 ps, which were used as starting structures to compute potential energy profiles of the hydroxylation reaction. The QM region comprised pOHB and the isoalloxazine part of FADHOOH up to the first methylene unit of the ribityl side chain, which was saturated with a hydrogen link atom. The B3LYP functional[76−78] in combination with a 6-31G* basis set was used to model the QM atoms. When applying the SMBP, the inner region was centered on the initial position of the distal oxygen atom of the FADHOOH cofactor with an extended dielectric cavity radius of 22.5 Å. All charge groups with any atom within 18.5 Å of the center were assigned to the inner region and were modeled explicitly. The remaining set of atoms constituted the outer macromolecule region and was represented by the SMBP. All charge groups with any atom within 16 Å of the center were assigned to the active region and were allowed to move during optimization. The hydroxylation reaction was described by means of a RC, which was chosen to be the difference in the bond lengths of the breaking bond between the proximal ($O_p$) and distal ($O_d$) oxygen atoms of FADHOOH and the forming bond

***Table 7.*** Reaction Energies ($\Delta E$) and Activation Energies ($\Delta E^{\ddagger}$) of the Hydroxylation Reaction in *p*-Hydroxybenzoate Hydroxylase [kcal/mol] Computed with Different Treatments of Long-Range Electrostatics

| | Coulomb | | SMBP(vac)[a] | | NOR[b] | | SMBP(solv)[c] | | SMBP(solv,app)[d] | |
|---|---|---|---|---|---|---|---|---|---|---|
| configuration | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ | $\Delta E$ | $\Delta E^{\ddagger}$ |
| 1 | −27.0 | 9.5 | −26.9 | 9.4 | −28.9 | 11.4 | −28.5 | 8.0 | −29.7 | 7.9 |
| 3 | −21.2 | 11.4 | −20.9 | 11.4 | −35.5 | 4.6 | −21.8 | 10.7 | −30.6 | 7.9 |
| 4 | −28.1 | 7.9 | −27.9 | 7.9 | −32.1 | 6.6 | −29.8 | 5.9 | −31.3 | 6.1 |
| 5 | −25.3 | 10.1 | −25.2 | 10.2 | −29.5 | 7.9 | −25.2 | 9.0 | −28.9 | 8.1 |
| mean value | −25.4 | 9.7 | −25.2 | 9.7 | −31.5 | 7.6 | −26.3 | 8.4 | −30.1 | 7.5 |
| standard deviation of data[e] | 3.0 | 1.5 | 3.1 | 1.5 | 3.0 | 2.9 | 3.6 | 2.0 | 1.0 | 1.0 |
| standard deviation of mean[f] | 1.5 | 0.7 | 1.5 | 0.7 | 1.5 | 1.4 | 1.8 | 1.0 | 0.5 | 0.5 |
| MAD[g] | – | – | 0.2 | 0.0 | 6.1 | 3.0 | – | – | 3.8 | 1.0 |

[a] $\varepsilon = 1$. [b] Neglect of outer region. [c] $\varepsilon = 80$. [d] $\varepsilon = 80$, $\phi_{rf}^{QM} = 0$. [e] Standard deviation of individual energy values. [f] Standard deviation of the mean value (68% confidence limit). [g] Mean absolute deviation relative to full Coulombic electrostatics. For SMBP(solv,app), SMBP(solv) values are used as a reference.

between the distal oxygen atom and the meta carbon atom of pOHB ($C_m$).

$$\xi = d(O_d - O_p) - d(O_d - C_m) \tag{7}$$

The potential energy profile was scanned for RC values between −1.7 Å and +1.7 Å in steps of 0.1 Å. We applied the same statistical test as in the CM system to validate the boundary potentials for PHBH and determined the optimal mesh and basis set sizes. For the sake of brevity and clarity, the data are relegated to the SI (Tables S4−S6, Figure S1), and only the results are briefly summarized here. Analysis of the deviations of the gradient components (Table S5) indicates that deviations in PHBH are very similar to those in CM. The relationship of accuracy and radial position is also similar: deviations are very small in the center of the inner region and increase significantly only for coarse mesh sizes and only at the boundary separating the inner and outer regions (Figure S1). Therefore, a mesh size combination of 0.25/1.25 Å excels again as a safe choice that reproduces full Coulombic electrostatics very accurately in all parts of the inner region. However, coarser mesh sizes also reproduce the electrostatic potential accurately everywhere except in close proximity to the boundary. As in CM, potential energy differences are less sensitive to the mesh sizes: computations with mesh size combinations ranging from 0.25/1.25 Å to 1.8/3.5 Å yield accurate results with MAX deviations that do not exceed 0.3 kcal/mol for all mesh sizes (Table S7). In GSBP calculations, basis sets of similar size are adequate for CM and PHBH (Table S6).

On the basis of these results, we conclude that SMBP and GSBP behave very similarly with respect to their accuracy and its dependence on the parameters for these two enzymes, despite the different nature of the two reactions and the different QM methods used (AM1 and B3LYP, respectively). This substantiates our expectation that not only the validation protocol but also the resulting parameters are transferable to other enzymatic systems. We suggest a mesh size combination of 0.25/1.25 Å with multipole moments up to order $L = 19$ as a safe choice for accurate calculations. As default options, we recommend a mesh size combination of 0.6/1.25 Å with multipole moments up to order $L = 10$ for efficient application of SMBP and GSBP. Since even coarser mesh sizes reproduce energy differences with only marginal deviations, we are confident that these values yield accurate results at reduced computational costs for general enzymatic systems.

A mesh size combination of 0.25/1.25 Å was applied to compute the potential energy differences of the hydroxylation step in PHBH with the SMBP in solution and in a vacuum, with full Coulombic electrostatics, and with the simple NOR approach. We found it unavoidable to discard configuration 2, since we failed to compute continuous energy profiles for this configuration despite many attempts. The results in Table 7 show that the SMBP reproduces standard QM/MM results with high accuracy. The mean values of the reaction and activation energy deviate by only 0.2 and 0.0 kcal/mol. MAD values of similar magnitude show that the results for the individual configurations also agree well. In contrast to CM, the EPOM has a significant effect on the energetics of the hydroxylation reaction in PHBH. If the electrostatic influence of the outer region is simply neglected (NOR), the mean value of the reaction energy changes by more than 6 kcal/mol from −25.4 to −31.5 kcal/mol. The activation energy is reduced by more than 2 kcal/mol from 9.7 to 7.6 kcal/mol. MAD values of 6.1 and 3.0 kcal/mol for reaction and activation energies show that the deviations are even larger for the individual configurations.

A more detailed analysis reveals that two different effects are responsible for the observed differences in the reaction energies. In configurations 1 and 3, the hydrogen bonding networks connecting the QM and MM region are different after geometry optimization with full Coulombic electrostatics and with the NOR approximation. By contrast, in configurations 4 and 5, the hydrogen bonding networks between the QM and MM region are very similar for both electrostatic treatments. In these two configurations, the differences in the reaction energies are dominated by the differences in the electrostatic QM−MM interaction energies, as shown in Table S8. Here, the electrostatic QM−MM interaction energy includes direct QM−MM interactions as well as the polarization effect of the MM point charges. These results suggest that with the NOR approximation the MM atoms are more flexible in adapting to the electrostatic potential of the QM region since they do not feel the EPOM. This effect becomes more pronounced when the QM region becomes more polar. During the hydroxylation reaction in PHBH, the less polar peroxide group separates into more

polar alcohol and alcoholate groups, which form hydrogen bonds with the neighboring MM residues. Therefore, the difference in the QM−MM interaction energies of the product and reactant is greater with the NOR approximation compared to that with full Coulombic electrostatics. The NOR approximation favors the more polar product state and thus shifts the reaction energies to more negative values. In conclusion, we face two possible consequences of neglecting the EPOM: either geometry optimizations may lead to structures that already differ in the hydrogen bonding network or changes in the QM−MM interaction energies may bias reaction energetics significantly, even when the hydrogen bonding network is similar. These results underline the fact that the EPOM can have a significant influence on potential energy differences. They cast doubt on the quantitative accuracy of QM/MM and pure QM studies that neglect the EPOM.

In PHBH, we also observe a significant solvent effect on the reaction. When bulk solvent is modeled with the SMBP, the activation energy decreases by more than 1 kcal/mol from 9.7 to 8.4 kcal/mol. This agrees qualitatively with chemical reasoning: the bulk solvent stabilizes the charged $OH^+$ species and lowers the energy of the transition state. Given the distinct charge transfer in this reaction, the effect of bulk solvent that is observed in PHBH should be in the upper range of what can be expected in enzymatic reactions. Even stronger effects seem possible if the inner region is chosen to be rather small and the charge transfer thus occurs closer to the dielectric boundary.

The significant solvent effect renders this reaction an interesting test for the SMBP(solv,app) method, which neglects the QM contribution to the reaction field potential. A satisfying agreement with SMBP(solv) results can only be observed for configuration 1. For the other configurations, deviations reach up to 10 kcal/mol. The mean value for the reaction energy is −30.1 kcal/mol and therefore far outside the confidence interval of the SMBP(solv) mean value. In the case of PHBH, the QM contribution to the reaction field potential thus has a significant influence on the energetics of the hydroxylation reaction. Hence, the SMBP(solv,app) method is not a valid approximation in the PHBH system.

## 5. Conclusion

In this study, we evaluated the electrostatic effect of the outer macromolecule region and of bulk solvent on two enzymatic systems. We applied the newly developed SMBP, which allows us to model bulk solvent as a PDC and to distinguish the effects of bulk solvent and the EPOM. The SMBP introduces approximations to describe electrostatic interactions with the outer macromolecule region more efficiently. Therefore, the accuracy of the SMBP was evaluated for both enzymatic test systems, and a protocol for validation and determination of adequate values for its inherent parameters was presented. This protocol was applied to generate a set of optimal parameters that is transferable to general enzymatic systems. The SMBP was found to describe the EPOM with high accuracy. Typically, deviations of mean values on the order of 0.1 to 0.2 kcal/mol are observed for both enzymes. Deviations for individual configurations may be slightly higher but rarely exceed 0.3 kcal/mol.

Two enzymatic reactions with rather different characteristics were used to study the effect of the EPOM and bulk solvent. The Claisen rearrangement in CM is a pericyclic reaction without much charge transfer. For this kind of reaction, the electrostatic influence of the outer macromolecule region on the reaction energetics is not significant when considering the mean values of all 10 configurations. However, deviations on the order of 1.5 kcal/mol are observed for individual configurations so that the neglect of long-range electrostatics can be detrimental in the absence of adequate sampling. Bulk solvent effects on the reaction energetics in CM are found to be small.

The hydroxylation reaction in PHBH, in contrast, is associated with a stronger charge transfer since the reaction formally corresponds to an $OH^+$ transfer. As a consequence, the EPOM has a strong influence on the reaction energetics, and its neglect causes errors of several kilocalories per mole due to a systematic overstabilization of the more polar product state (arising from the higher flexibility of the MM residues without the EPOM). Moreover, bulk solvent stabilizes the transition state and reduces the reaction barrier by about 1 kcal/mol in PHBH.

Depending on the charge transfer characteristics of the chemical process, the EPOM and bulk solvent can thus have a significant effect on the energetics of enzymatic reactions. Among these two contributions, the EPOM is clearly more important. The SMBP offers a convenient way to evaluate both contributions accurately and efficiently in QM/MM calculations.

**Acknowledgment.** This work was supported by the Max Planck Initiative on Multiscale Materials Modeling and the Volkswagenstiftung. T.B. gratefully acknowledges a Kekulé scholarship from the Fonds der Chemischen Industrie. We thank J. Breidung and S. Thiel for CM and PHBH input data and A. Koslowski for his program to perform spline fits of energy profiles.

**Supporting Information Available:** Energy differences of the reactions in CM and PHBH with different mesh size combinations; free energy differences of the reaction in CM using full Coulombic electrostatics for two configurations; evaluation of the accuracy of the electrostatic forces for different basis set sizes using the GSBP for CM and PHBH; evaluation of the accuracy of the electrostatic forces with different mesh sizes for PHBH; contributions to the reaction energies of the hydroxylation reaction in PHBH; plot of MAD and MAX values of electrostatic forces as a function of the radial position for PHBH. This information is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Struct.* **1999**, *28*, 155–179.

(2) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211–217.

(3) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509–521.

(4) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186–195.

(5) Monticelli, L.; Simãţes, C.; Belvisi, L.; Colombo, G. *J. Phys.: Condens. Matter* **2006**, *18*, S329–S345.

(6) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–443.

(7) Schreiber, H.; Steinhauser, O. *Biochemistry* **1992**, *31*, 5856–5860.

(8) Schreiber, H.; Steinhauser, O. *Chem. Phys.* **1992**, *168*, 75–89.

(9) Baumketner, A.; Shea, J.-E. *J. Phys. Chem. B* **2005**, *109*, 21322–21328.

(10) Patra, M.; Karttunen, M.; Hyvönen, M.; Falck, E.; Vattulainen, I. *J. Phys. Chem. B* **2004**, *108*, 4485–4494.

(11) Patra, M.; Karttunen, M.; Hyvönen, M.; Falck, E.; Lindqvist, P.; Vattulainen, I. *Biophys. J.* **2003**, *84*, 3636–3645.

(12) Ewald, P. *Ann. Phys.* **1921**, *369*, 253–287.

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

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

(15) Hünenberger, P. H.; McCammon, J. A. *Biophys. Chem.* **1999**, *78*, 69–88.

(16) Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856–1872.

(17) Kuwajima, S.; Warshel, A. *J. Chem. Phys.* **1988**, *89*, 3751–3759.

(18) Weber, W.; Hünenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668–3675.

(19) Esselink, K. *Comput. Phys. Commun.* **1995**, *87*, 375–395.

(20) Appel, A. W. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 85–103.

(21) Schmidt, K. E.; Lee, M. A. *J. Stat. Phys.* **1991**, *63*, 1223–1235.

(22) Ding, H. Q.; Karasawa, N.; Goddard, W. A., III. *J. Chem. Phys.* **1992**, *97*, 4309–4315.

(23) Berkowitz, M.; McCammon, J. A. *Chem. Phys. Lett.* **1982**, *90*, 215–217.

(24) Brooks, C. L., III; Karplus, M. *J. Chem. Phys.* **1983**, *79*, 6312–6325.

(25) Brunger, A.; Brooks, C. L., III; Karplus, M. *Chem. Phys. Lett.* **1984**, *105*, 495–500.

(26) Essex, J. W.; Jorgensen, W. L. *J. Comput. Chem.* **1995**, *16*, 951–972.

(27) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.

(28) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(29) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2–13.

(30) Gao, J.; Alhambra, C. *J. Chem. Phys.* **1997**, *107*, 1212–1217.

(31) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019–1031.

(32) Yarne, D. A.; Tuckerman, M. E.; Martyna, G. J. *J. Chem. Phys.* **2001**, *115*, 3531–3539.

(33) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1*, 1176–1184.

(34) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2006**, *2*, 1370–1378.

(35) Friedman, H. L. *Mol. Phys.* **1975**, *29*, 1533–1543.

(36) Wang, L.; Hermans, J. *J. Phys. Chem.* **1995**, *99*, 12001–12007.

(37) Brunger, A.; Brooks, C. L., III; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 8458–8462.

(38) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100–3107.

(39) Alper, H.; Levy, R. M. *J. Chem. Phys.* **1993**, *99*, 9847–9852.

(40) Warshel, A.; King, G. *Chem. Phys. Lett.* **1985**, *121*, 124–129.

(41) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(42) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.

(43) Im, W.; Bernèche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924–2937.

(44) Banavali, N. K.; Im, W.; Roux, B. *J. Chem. Phys.* **2002**, *117*, 7381–7388.

(45) König, P. H.; Ghosh, N.; Hoffmann, M.; Elstner, M.; Tajkhorshid, E.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. A* **2006**, *110*, 548–563.

(46) Ma, L.; Cui, Q. *J. Am. Chem. Soc.* **2007**, *129*, 10261–10268.

(47) Zhu, X.; Jethiray, A.; Cui, Q. *J. Chem. Theory Comput.* **2007**, *3*, 1538–1549.

(48) Riccardi, D.; König, P.; Prat-Resina, X.; Yu, H.; Elstner, M.; Frauenheim, T.; Cui, Q. *J. Am. Chem. Soc.* **2006**, *128*, 16302–16311.

(49) Schaefer, P.; Riccardi, D.; Cui, Q. *J. Chem. Phys.* **2005**, *123*, 014905/1−14.

(50) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2008**, *4*, 1600–1609.

(51) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 3114–3128.

(52) Senn, H. M.; Kästner, J.; Breidung, J.; Thiel, W. *Can. J. Chem.* **2009**, *87*, 1322–1337.

(53) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *56*, 467–505.

(54) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 2925–2929.

(55) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

(56) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483–3492.

(57) Sherwood, P.; et al. *THEOCHEM* **2003**, *632*, 1–28.

(58) ChemShell. http://www.chemshell.org (accessed August 14, 2010).

(59) Thiel, W. *MNDO2004*; Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2004.

(60) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(61) Smith, W.; Forester, T. *J. Mol. Graph.* **1996**, *14*, 136–141.

(62) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(63) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177–2186.

(64) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterlig, W. T. Partial Differential Equations. In *Numerical Recipes in C*; Cambridge University Press: Cambridge, England, 1988; pp 673−680.

(65) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(66) Im, W.; Seefeld, S.; Roux, B. *Biophys. J.* **2000**, *79*, 788–801.

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

(68) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

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

(70) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(71) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(72) Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, T.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(73) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(74) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1987**, *9*, 327–335.

(75) Kast, P.; Aisf-Ullah, M.; Hilvert, D. *Tetrahedron Lett.* **1996**, *37*, 2691–2694.

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

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

(78) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(79) Senn, H. M.; Thiel, S.; Thiel, W. *J. Chem. Theory Comput.* **2005**, *1*, 494–505.

(80) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H.-J. *Angew. Chem.* **2006**, *45*, 6856–6859.

(81) Mata, R. A.; Werner, H.-J.; Thiel, S.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 025104/1−8.

CT1005455