JCTC Journal of Chemical Theory and Computation

# Folding Peptides into Lipid Bilayer Membranes

Martin B. Ulmschneider* and Jakob P. Ulmschneider*

*Department of Biochemistry, University of Oxford, Oxford OX1 1QU, U.K.*

Received March 24, 2008

**Abstract:** The adsorption, insertion, and folding of a synthetic 16-residue WALP peptide was studied at physiological time scales (>$\mu$s) by atomic detail molecular dynamics simulation using a fully explicit DPPC/DMPC lipid bilayer setup. The temperature was elevated to 80 °C/44 °C respectively to increase sampling. After spontaneous adsorption the peptide crosses the polar interfaces to locate at the hydrophobic bilayer core. Remarkably, insertion occurs before folding, and the dominant configurations are inserted beta-hairpins. For the DPPC simulation a trans-membrane helix formed but was not stable. Unfolded membrane insertion of WALP was first observed by Nymeyer and co-workers using a replica exchange method. However, both results are in stark contrast to current theory and simulations with implicit membrane models, which rule out unfolded insertion into the hydrophobic core. At present the exact reasons for this unexpected behavior cannot be unambiguously determined, due to the lack suitable experimental and simulation data to compare to. Nevertheless, the results demonstrate that simulation studies can now in principle provide atomic detail insights into complex biophysical phenomena at physiologically relevant time scales. Future effort must now concentrate on suitable ways to verify current force fields and methodologies for such simulations.

Folding and integration of peptides into lipid bilayer membranes remains one of the most intriguing processes in biophysics, as it cannot be directly observed at sufficient temporal and spatial resolutions. Recent experiments applying the translocon machinery to insert designed peptides[1,2] as well as statistical analyses of membrane protein structures[3] revealed that the distributions of individual amino acid types correlate strongly with their expected solvation energies along the membrane normal.[4] In addition, the Sec translocon structure revealed that nascent peptides are threaded into a narrow water filled channel, which opens laterally allowing hydrophobic segments to partition into the bilayer.[5] Many details of this process, including how much folding actually occurs inside the channel are currently unclear, but direct peptide-bilayer interactions (i.e., the solvation free energy) seem to be the key determining the partitioning and folding properties of a particular sequence. From a physical chemistry perspective transfer of solvated peptides into a hydrocarbon phase should follow a two-stage pathway,[6] where helical segments fold at the phase boundary before inserting, due to the high cost (estimated at ~4 kcal/mol) of desolvating exposed peptide bonds.[7,8]

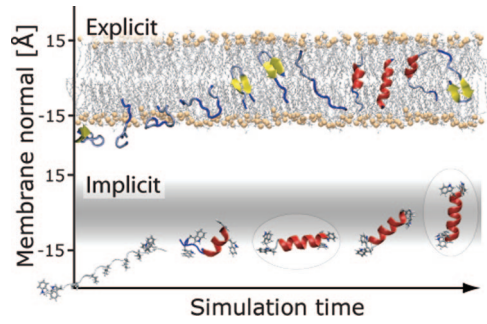



***Figure 1.*** Folding pathways of the WALP peptide observed in an explicit lipid bilayer and the implicit membrane at 80 °C. The pathways is markedly different. In the implicit membrane, based on generalized Born theory, the folding pathway follows the two-state model as expected from current partitioning theory (adsorption, interfacial folding and folded insertion). In the explicit system there is no interfacial state, and insertion occurs in an unfolded conformation.

Computer simulations using implicit solvent models,[9,10] which treat the membrane as a low-dielectric slab, have confirmed this folding pathway for simple hydrophobic peptides,[11,12] including the synthetic WALP peptide (see Figure 1).[13] The peptide was found to quickly adsorb to the membrane surface. Stable insertion was only observed after interfacial folding into helical conformers. However, these models lack the complex polar bilayer interfaces and neglect entropic effects due to the liquid crystalline order of the lipids and water molecules and therefore do not represent a realistic membrane.

We therefore repeated the simulation in a fully explicit setup using the GROMOS96 force field (methods are given in the Supporting Information). As in the implicit system WALP was placed into bulk water and positioned parallel to the surface of a DPPC lipid bilayer in a completely extended conformation. The temperature was elevated to 80 °C to increase sampling

* Corresponding author e-mail: martin@ulmschneider.com (M.B.U.); jakob@ulmschneider.com (J.P.U.).

10.1021/ct800100m CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/03/2008

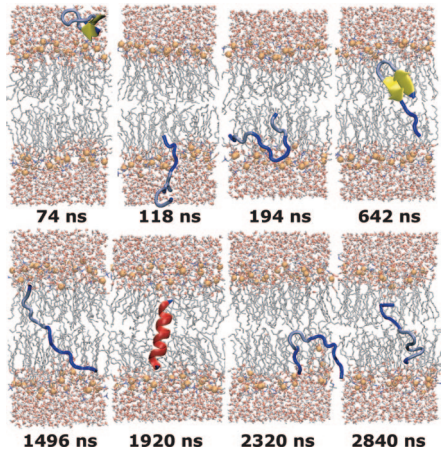



***Figure 2.*** Adsorption, insertion, and folding of the WALP peptide in an explicit DPPC/SPC bilayer system. The peptide passes the lipid head-groups in an unfolded conformation and inserts into the hydrophobic core before folding. A large number of beta-structures are sampled, and several water molecules are dragged into the hydrocarbon core of the membrane.

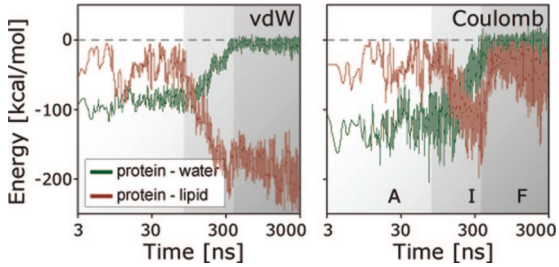



***Figure 3.*** Interaction energies of the peptide with the water and lipid bilayer. Favorable van der Waals interactions of the peptide with the lipids are the main driving force for insertion, which occurs in the first 500 ns. Time is plotted logarithmically indicating the three distinctive phases: adsorption (**A**), insertion (**I**), and folding (**F**).

and ensure the lipids remain in the fluid phase. Below this temperature stably inserted WALP helices were observed to be the dominant configuration in the implicit membrane.[12]

The simulation was run for 3 $\mu$s and shows strikingly unexpected behavior (see Figure 2). First, adsorption is slow, and the peptide remains fully solvated for ~100 ns. After precipitating onto the membrane surface the peptide immediately starts to cross the phosphocholine head-groups. This process is driven by favorable van der Waals interactions and takes ~300 ns. During this time the peptide remains virtually completely unfolded, and lost electrostatic interactions with the water molecules are compensated by interactions with the polar head-groups, which are of the same order of magnitude (see Figure 3). Subsequently the peptide inserts into the hydrophobic membrane core in an unfolded configuration, and for the remaining ~2 $\mu$s the peptide oscillates between deeply inserted completely unfolded and misfolded conformations (beta-hairpins). Both the absence of a stable interfacial state and the unfolded insertion are remarkable, as is the persistence of fully inserted beta-hairpins, which are the dominant configuration (Figure 4). Such conformations should be less favorable than helices since only half the backbone hydrogen bonds can be

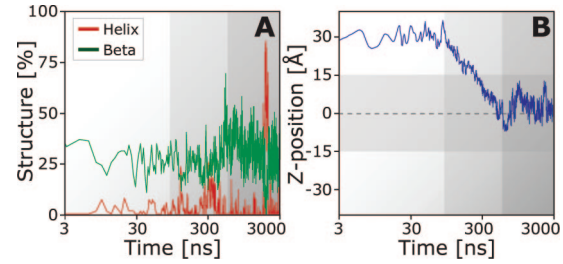



***Figure 4.*** Peptide secondary structure (**A**) and insertion depth (**B**). The membrane center is at $z = 0$, and the hydrophobic core is shaded.

satisfied. Eventually, after ~1.9 $\mu$s rapid formation of a helical conformer is observed from a completely extended membrane spanning configuration. Surprisingly, however, this is not concomitant with further energetic stabilization as would be expected, and the helix remains stable for only ~200 ns before unfolding again.

An unfolded insertion pathway for WALP was first observed in explicit bilayer replica exchange simulations by Nymeyer and co-workers, on a time scale of ~4 ns, with the CHARMM22 force field.[14] Similar to the results presented here, insertion of unfolded conformations preceded the formation of helical structure. However, contrary to our results the trans-membrane helix was found to be stable at 350 K, the lowest temperature replica. Instead, we observe only marginal stability at roughly the same temperature (353 K), with several folding/unfolding events and no clear energetic stabilization of helical conformations over unfolded ones. In a control simulation an inserted helix also unfolded after ~300 ns, indicating that it is not stable at 80 °C.

We performed several other control simulations. A 1 $\mu$s simulation of WALP in a DMPC bilayer, which has a lower chain melting temperature of ~25 °C also showed unfolded insertion. However, unlike the DPPC case the peptide inserts only in one leaflet of the bilayer, and insertion is generally less deep than in the DPPC case. Due to the lower simulation temperature of 44 °C sampling was found to be significantly slower, which might explain the absence of trans-membrane configurations. The insertion process took ~400 ns, and no helical folding was observed.

We also performed a 1 $\mu$s simulation of an unfolded, inserted conformer in DPPC at a reduced temperature of 50 °C. The peptide was found to remain stably inserted in an unfolded configuration. Sampling is severely reduced, and some ordering of the lipids is observed as expected for temperatures close to the DPPC chain melting temperature.

At room temperature WALP generally forms stable membrane spanning helices. Unfortunately, no structural data are currently available for WALP at elevated temperatures. Furthermore, the insertion process is very difficult to study experimentally since few methods can provide the necessary submicrosecond time-resolution. Gai and co-workers have recently reported a promising method to explore the coil to helix transition associated with membrane binding via fluorescence resonance energy transfer.[15] This method has been applied to a range of peptides, and while they generally find folding to occur before insertion, they do not rule out unfolded insertion.

It should be noted that the experimental internalization of WALP into bilayers might differ from the present simulations. Like most hydrophobic sequences WALP is not soluble and aggregates in water. Insertion is therefore achieved by hydration of a WALP-lipid suspension, resulting in vesicles containing the peptide in a membrane spanning conformation, generally assumed to be helical.[16] Simulation of membrane peptides generally describe 'infinite' dilution scenarios (i.e., one WALP per bilayer), which is a reasonable approximation at low peptide-lipid ratios.

The above results can be interpreted in three ways: i. Unfolded insertion is an artifact and highlights major force field issues (water, lipid, peptide). This could be due to overestimation of the van der Waals interactions and entropic effects and/or an underestimation of the Coulomb terms and hence the solvation energies. This interpretation is supported by currently established theory, which dictates that only helical conformers can insert and reside stably in the membrane, due to the huge ~4 kcal/mol penalty to break a backbone hydrogen bond in the membrane.[8] ii. Folding and insertion of membrane bound peptides varies strongly with subtle temperature differences. iii. The results are genuine and suggest that entropic terms can compensate for the huge desolvation penalties associated with unfolded inserted conformers. This view can be supported by previous simulations of Nymeyer et al. who also observe unfolded insertion and subsequent folding for WALP using a different force field.[14]

Recent computational studies have revealed that the contributions of alanine and leucine side chain analogs, which make up the core of WALP residues, are favorable,[17−19] as expected from statistical analyses of membrane proteins and translocon mediated peptide insertion experiments.[1−3,20] However, burial of the peptide backbone as well as the formation of internal hydrogen bonds greatly contributes to the overall partitioning properties of peptides. In the absence of solid experimental evidence it is very difficult to conclude which of the many contributions is the dominant one since peptide partitioning into lipid bilayers is a complex process that requires very long time scale simulations with a large number of adjustable parameters (choice of lipids, temperature, force field parameters, simulation algorithms, etc.). More computational and experimental studies are needed to ultimately clarify these issues.

Since the simulation protocol and force field used are being very extensively employed by a large number of groups,[21] the results raise several important issues regarding long time scale simulations. The most important of which are the fundamental questions of how such simulations are to be verified and if current force field parameters and simulation methods are suited to allow accurate predictions of complex systems at physiological time scales.

**Acknowledgment**. We thank Prof. M. S. P. Sansom for support and use of facilities.

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

## References

(1) Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. *Nature* **2005**, *433* (7024), 377.

(2) Hessa, T.; Meindl-Beinker, N. M.; Bernsel, A.; Kim, H.; Sato, Y.; Lerch-Bader, M.; Nilsson, I.; White, S. H.; von Heijne, G. *Nature* **2007**, *450* (7172), 1026.

(3) Ulmschneider, M. B.; Sansom, M. S.; Di Nola, A. *Proteins* **2005**, *59* (2), 252.

(4) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27* (5), 1664.

(5) Van den Berg, B.; Clemons, W. M., Jr.; Collinson, I.; Modis, Y.; Hartmann, E.; Harrison, S. C.; Rapoport, T. A. *Nature* **2004**, *427* (6969), 36.

(6) Popot, J. L.; Engelman, D. M. *Biochemistry* **1990**, *29* (17), 4031.

(7) White, S. H.; Ladokhin, A. S.; Jayasinghe, S.; Hristova, K. *J. Biol. Chem.* **2001**, *276* (35), 32395.

(8) White, S. H. *Adv. Protein Chem.* **2006**, *72*, 157.

(9) Im, W.; Feig, M.; Brooks, C. L., III *Biophys. J.* **2003**, *85*, 2900.

(10) Ulmschneider, M. B.; Ulmschneider, J. P.; Sansom, M. S. P.; Di Nola, A. *Biophys. J.* **2007**, *92*, 2338.

(11) Im, W.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6771.

(12) Ulmschneider, M. B.; Ulmschneider, J. P. *Mol. Membr. Biol.* **2008**, *25* (3), 245.

(13) Killian, J. A. *FEBS Lett.* **2003**, *555* (1), 134.

(14) Nymeyer, H.; Woolf, T. B.; Garcia, A. E. *Proteins* **2005**, *59* (4), 783.

(15) Tucker, M. J.; Tang, J.; Gai, F. *J. Phys. Chem. B* **2006**, *110* (15), 8105.

(16) de Planque, M. R. R.; Goormaghtigh, E.; Greathouse, D. V.; Koeppe, R. E.; Kruijtzer, J. A. W.; Liskamp, R. M. J.; de Kruijff, B.; Killian, J. A. *Biochemistry* **2001**, *40* (16), 5000.

(17) Johansson, A. C.; Lindahl, E. *Proteins* **2008**, *70* (4), 1332.

(18) MacCallum, J. L.; Bennett, W. F.; Tieleman, D. P. *Biophys. J.* **2008**, *94* (9), 3393.

(19) Li, L.; Vorobyov, I.; Allen, T. W. *J. Phys. Chem. B* **2008**, *112* (32), 9574.

(20) Ulmschneider, M. B.; Sansom, M. S. P. *Biochim. Biophys. Acta* **2001**, *1512* (1), 1.

(21) Efremov, R. G.; Nolde, D. E.; Konshina, A. G.; Syrtcev, N. P.; Arseniev, A. S. *Curr. Med. Chem.* **2004**, *11* (18), 2421.

CT800100M

JCTC Journal of Chemical Theory and Computation

# Degenerate Perturbation Theory for Electronic g Tensors: Leading-Order Relativistic Effects

Zilvinas Rinkevicius,* Katia Julia de Almeida, Cornel I. Oprea, Olav Vahtras, and Hans Ågren

*Department of Theoretical Chemistry, School of Biotechnology, Royal Institute of Technology, SE-10691 Stockholm, Sweden*

Kenneth Ruud

*Centre for Theoretical and Computational Chemistry, Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway*

Received February 20, 2008

**Abstract:** A new approach for the evaluation of the leading-order relativistic corrections to the electronic g tensors of molecules with a doublet ground state is presented. The methodology is based on degenerate perturbation theory and includes all relevant contributions to the g tensor shift up to order $\mathcal{O}(\alpha^4)$ originating from the one-electron part of the Breit−Pauli Hamiltonian—that is, it allows for the treatment of scalar relativistic, spin−orbit, and mixed corrections to the spin and orbital Zeeman effects. This approach has been implemented in the framework of spin-restricted density functional theory and is in the present paper, as a first illustration of the theory, applied to study relativistic effects on electronic g tensors of dihalogen anion radicals $X_2^-$ (X = F, Cl, Br, I). The results indicate that the spin−orbit interaction is responsible for the large parallel component of the g tensor shift of $Br_2^-$ and $I_2^-$, and furthermore that both the leading-order scalar relativistic and spin−orbit corrections are of minor importance for the perpendicular component of the g tensor in these molecules since they effectively cancel each other. In addition to investigating the g tensors of dihalogen anion radicals, we also critically examine the importance of various relativistic corrections to the electronic g tensor of linear molecules with Σ-type ground states and present a two-state model suitable for an approximate estimation of the g tensor in such molecules.

## I. Introduction

Electron paramagnetic resonance (EPR) spectroscopy is an important experimental technique designed for the investigation of compounds with unpaired electrons. The information content of this technique is obtained by measurements of molecular parameters related to the Zeeman effect—that is, the electronic g tensor, the hyperfine coupling constants, and the zero-field splitting constants.[1−3] These parameters play a central role in the interpretation of EPR measurements and are often connected to the electronic and geometrical structure of the systems analyzed using empirical relationships.[2,3] This holds in particular for classes of species with well-established empirical "molecular structure—Zeeman effect parameter" relationships, while for more complex systems, the empirical approach often encounters difficulties and can lead to multiple interpretations of the experimental results. One way to enhance the quality of the interpretation of EPR spectra and to overcome the ambiguity introduced by the use of empirical relationships is to incorporate quantum chemical modeling of the molecular parameters into the analysis.[2,3] This combined approach for extracting information from EPR measurements is already now commonly in use[2,3] and is expected to become even more prominent with the improved applicability of quantum chemistry methods to larger molecular systems of experimental interest.

10.1021/ct800053f CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/28/2008

A multitude of approaches for evaluating electronic g tensors of molecules has been developed and presented in the literature in recent years.[4−20] These methods range from sophisticated ab initio approaches such as multiconfigurational self-consistent field (MCSCF) response theory[6,17] to various density functional theory (DFT) based formalisms.[7,8,10,12−14,16,18,19] The accuracy and applicability of these methodologies have been intensively studied and recipes for accurate calculations of electronic g tensors of organic radicals and first-row transition metal complexes have been suggested in several recent publications.[10,12,14,16] Despite these achievements, only a few works[8,9,13,18−26] have targeted compounds containing heavy elements, and relativistic effects on electronic g tensors have been studied in detail primarily for diatomics and other small compounds, such as the uranium hexafluoride anion,[21] to mention one prominent example. In the domain of density functional theory, a few attempts have been made at including relativistic effects in the evaluation of the electronic g tensors. Following the pioneering work of van Lenthe et al.,[8] the majority of DFT methods[8,13] capable of computing relativistic g tensors have been based on the two-component Kohn−Sham approach, in which the reduced two-component Hamiltonian is obtained by applying the zeroth-order regular approximation (ZORA)[27−30] or the Douglas−Kroll (DK)[31] transformation. The first implementations of two-component DFT methods for the evaluation of electronic g tensors were restricted to molecules with doublet ground states, since Kramer's doublet symmetry could then be utilized in the determination of the electronic g tensor components. This restriction was lifted only recently, when Malkin et al.[18] implemented a spin-polarized Douglas−Kroll Kohn−Sham formalism which can handle molecules with arbitrary ground states. Despite the obvious advantages of two-component electronic structure methods, these methods have until now only been applied to systems consisting of a few atoms. The only exception from this unfortunate trend is van Lenthe's ZORA formalism,[8] which allows for the rapid evaluation of the electronic g tensors in molecules with doublet ground states, and this approach has been applied many times in studies of compounds containing heavy elements.[32−34]

An alternative to the two-component DFT approaches is a one-component formalism in which the relativistic corrections to the electronic g tensor are treated on an equal footing to the nonrelativistic g tensor itself—that is, via perturbation theory expansions. Here, the term "non-relativistic g tensor" is understood as the g tensor composed of the free-electron $g$ factor, $g_e \approx 2.0023$, and g tensor shifts of order $\mathcal{O}(\alpha^2)$, i.e., $\vec{\vec{g}}_{NR} = g_e\vec{\vec{1}} + \Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$. Consequently, the leading-order relativistic corrections to the g tensor are contributions to the g tensor shift of order $\mathcal{O}(\alpha^4)$, where $\alpha$ is the fine-structure constant, which is approximately 1/137 in the atomic unit system used throughout this paper. However, this approach has been exploited only at the ab initio theory level[17] due to the lack of density functional theory implementations capable of handling arbitrary perturbations beyond second order. We recently extended our spin-restricted open-shell density functional response theory from the linear[35] to the quadratic level,[36] and this opens the way for computing the leading-order relativistic corrections to the electronic g tensors using perturbation theory at the DFT level for the first time. In this paper, we apply degenerate perturbation theory (DGPT) for molecules with a doublet ground state to derive formulas for evaluating electronic g tensors in which relativistic corrections are accounted for, and employ linear and quadratic response functions as well as their residues to determine the different contributions to the g tensor appearing in the formalism at the level of spin-restricted density functional theory. The computational procedure proposed in this work extends previous simplified treatments of relativistic corrections to the electronic g tensor developed by Manninen, Vaara, and Ruud,[17] and corrects conceptually as well as computationally deficiencies in that formalism. Apart from describing the formalism, we also present the results of electronic g tensor calculations for the series of dihalogen anion radicals $X_2^-$ (X = F, Cl, Br, I). In addition to testing the performance of the perturbational DFT treatment of relativistic electronic g tensors, we also resolve contradictions between the works of Manninen et al.[17] and Malkin et al.[18] regarding the dominant physical mechanism responsible for the large parallel g tensor shift of the dihalogen anion radicals.

The paper is organized in the following way. First, we give a detailed description of the relativistic electronic g tensor methodology based on degenerate perturbation theory, and we subsequently describe the computational details and discuss the results obtained for the dihalogen anion radicals. Finally, we draw some conclusions and outline some prospects for further applications of the formalism presented here.

## II. Theory

Experimentally, observed EPR spectra are usually analyzed using a phenomenological spin Hamiltonian approach[1−3] in which the Zeeman effect is described as

$$\hat{\mathcal{H}}_{SH} = \mu_B \vec{B}\cdot\vec{\vec{g}}\cdot\vec{\mathcal{S}} \tag{1}$$

where $\mu_B = 1/2$ is the Bohr magneton and $\vec{\vec{g}}$ is the electronic g tensor of the molecule, which mediates the interaction between the effective spin $\vec{\mathcal{S}}$ and a static external magnetic field $\vec{B}$. In the special case of a molecule with a single unpaired electron, the effective spin $\vec{\mathcal{S}}$ can be rewritten in terms of the Pauli matrices $\vec{\mathcal{S}} = 1/2\vec{\boldsymbol{\sigma}}$, and the spin Hamiltonian becomes a $2 \times 2$ matrix

$$\hat{\mathcal{H}}_{SH} = \frac{\mu_B}{2}\vec{B}\cdot\vec{\vec{g}}\cdot\vec{\boldsymbol{\sigma}} \tag{2}$$

Consequently, the eigenvalues of $\hat{\mathcal{H}}_{SH}$ can readily be determined in a basis of two-dimensional spinors, leading to the well-known expression of Abraham and Bleaney[37] for the Zeeman splitting

$$\Delta E = \mu_B\sqrt{\vec{B}\cdot\vec{\vec{G}}\cdot\vec{B}} \quad \text{and} \quad \vec{\vec{G}} = \vec{\vec{g}}\cdot\vec{\vec{g}}^T \tag{3}$$

where we have introduced the so-called symmetric G tensor $\vec{\vec{G}}$, which is measured in typical EPR experiments. Despite the fact that the G tensor is directly obtained from experi-

ment, the theoretical and experimental results are usually analyzed and discussed in terms of the electronic g tensor $\vec{\vec{g}}$ which, being defined via three g factors with corresponding principal axes, is a more convenient quantity for interpretation. The electronic g tensor can be computed from the G tensor according to eq 3 in a two-step procedure: the first step is the diagonalization of the $\vec{\vec{G}}$ tensor, which determines the principal values and axes of this tensor; the second step involves the computation of the three g factors (diagonal components of the g tensor) as the positive square root of the principal values of the G tensor followed by an appropriate assignment of the principal axis to the $\vec{\vec{g}}$ components. This procedure is not only used to extract the electronic g tensor from experimental data, but is also employed in theoretical calculations for transforming the nonsymmetric $\vec{\vec{g}}$ into its diagonal form. Therefore, both the $\vec{\vec{G}}$ and $\vec{\vec{g}}$ tensors are equally suitable for a description of the Zeeman effect, and the selection of the specific tensor form depends only on convenience in the particular situation, as the G tensor always can be reduced to the electronic g tensor and vice versa, according to the procedure described above.

Before proceeding with the derivation of formulas for the evaluation of the relativistic g tensors, we briefly outline the connection between the spin Hamiltonian and the quantum mechanical treatment of the Zeeman effect in systems with ground states that in the absence of a magnetic field are described by a Kramers doublet pair $|\mathbf{0}\rangle = (|0\rangle|\bar{0}\rangle)$. In this case, the Zeeman interaction Hamiltonian is a $2 \times 2$ matrix

$$\hat{\mathbf{H}}_Z = \mu_B(\vec{\mathbf{L}} + 2\vec{\mathbf{S}})\cdot\vec{B} \tag{4}$$

where the angular momentum $\vec{\mathbf{L}}$ and electronic spin $\vec{\mathbf{S}}$ can be expanded in the basis of Pauli matrices

$$\vec{\mathbf{L}} = \frac{1}{2}\vec{\vec{\Lambda}}\cdot\vec{\boldsymbol{\sigma}} \quad \text{and} \quad \vec{\mathbf{S}} = \frac{1}{2}\vec{\vec{\Sigma}}\cdot\vec{\boldsymbol{\sigma}} \tag{5}$$

We observe that the Pauli matrices together with the unit matrix form a complete basis in the linear space of $2 \times 2$ matrices; for traceless operators, the expansion in eq 5 is the most general one. Taking into account the properties of the Zeeman interaction operator and the Kramers doublet pair, the Zeeman splitting can be written as the difference between the eigenvalues of $\hat{\mathbf{H}}_Z$[20]

$$\Delta E = \mu_B\sqrt{2\vec{B}\cdot\mathrm{tr}((\vec{\mathbf{L}} + 2\vec{\mathbf{S}})(\vec{\mathbf{L}} + 2\vec{\mathbf{S}}))\cdot\vec{B}} \tag{6}$$

Comparing this definition of the Zeeman splitting with the one obtained in the spin-Hamiltonian approach (see eq 3) and requiring that the equality between the left-hand sides of both equations is fulfilled for all strengths of the magnetic field $\vec{B}$, we obtain the following expression for calculating the G tensor

$$\vec{\vec{G}} = (\vec{\vec{\Lambda}} + 2\vec{\vec{\Sigma}})\cdot(\vec{\vec{\Lambda}} + 2\vec{\vec{\Sigma}})^T \tag{7}$$

which, according to eq 3, allows us to define the electronic g tensor as

$$\vec{\vec{g}} = \vec{\vec{\Lambda}} + 2\vec{\vec{\Sigma}} \tag{8}$$

From these equations it is evident that the $\vec{\vec{G}}$ and $\vec{\vec{g}}$ tensors are defined solely by the expansion tensors of the angular momentum $\vec{\mathbf{L}}$ and electronic spin $\vec{\mathbf{S}}$ operators, $\vec{\vec{\Lambda}}$ and $\vec{\vec{\Sigma}}$, and the task of computing the electronic g tensor reduces to the determination of the $\vec{\vec{\Lambda}}$ and $\vec{\vec{\Sigma}}$ tensors for a given approximate Hamiltonian using a selected ab initio or density functional theory method.

In this paper, we apply degenerate perturbation theory to the Kramers doublet pair in order to obtain formulas for evaluating the $\vec{\vec{\Lambda}}$ and $\vec{\vec{\Sigma}}$ tensors. The starting point for our derivation is the construction of the Kramers doublet pair $|\mathbf{0}\rangle$ as a row matrix from two degenerate doublet wave functions, $|^2 0;^1/_2\rangle$ and $|^2 0;-^1/_2\rangle$, which are eigenfunctions of the ordinary nonrelativistic Hamiltonian $\hat{H}_0$, the total spin $\hat{S}^2$ and the spin projection $\hat{S}_z$ operators, which satisfy the set of equations

$$\begin{aligned}
\hat{H}_0|^{2S+1}0;m\rangle &= E_0|^{2S+1}0;m\rangle \\
\hat{S}^2|^{2S+1}0;m\rangle &= S(S+1)|^{2S+1}0;m\rangle \\
\hat{S}_z|^{2S+1}0;m\rangle &= m|^{2S+1}0;m\rangle
\end{aligned}$$

where for a doublet state $S = 1/2$ and $m = 1/2, -1/2$, respectively. We here consider all one-electron spin- and external field-dependent operators as well as other relevant one-electron operators of order $\mathcal{O}(\alpha^2)$ in the Breit−Pauli Hamiltonian as perturbations. These operators were also considered as perturbations by Manninen et al.,[17] but we include in addition also the two-electron part of the spin−orbit operator, which was omitted in ref 17. Using these operators and their combinations, we will determine corrections to the Zeeman interaction eq 4 to $\mathcal{O}(\alpha^4)$ by applying degenerate perturbation theory (a brief account of this formalism is given in Appendix A).

At the nonrelativistic limit, i.e. order $\mathcal{O}(\alpha^0)$, the Zeeman effect is described by the spin and orbital Zeeman effect operators only (see Table 1, in which a detailed description of the operators included in the perturbation operator $\hat{V}$ is given), and for an orbitally nondegenerate doublet state we thus obtain the following contributions to the Zeeman effect from first-order DGPT

$$\mathbf{H}_{SZ} = \langle\mathbf{0}|\hat{H}_{SZ}|\mathbf{0}\rangle = \mu_B \sum_{k=x,y,z} B_k[\Sigma_{SZ}]_{kk}\boldsymbol{\sigma}_k = \mu_B \sum_{k=x,y,z} B_k\boldsymbol{\sigma}_k \tag{9}$$

$$\mathbf{H}_{OZ} = \langle\mathbf{0}|\hat{H}_{OZ}|\mathbf{0}\rangle = \frac{\mu_B}{2} \sum_{k=x,y,z} B_k[\Lambda_{OZ}]_{kk}\boldsymbol{\sigma}_k = 0 \tag{10}$$

which indicates that the magnetic moment of the molecule is defined by the spin of the unpaired electron in the molecule only. Therefore, at this level of approximation, the g tensor is $\vec{\vec{g}} = 2\vec{\vec{\Sigma}}_{SZ} = 2\vec{\vec{1}}$, and the molecule behaves as a free electron in a magnetic field. After adding the quantum electrodynamic correction to the magnetic moment of the electron we retrieve the well-known form of the g tensor of the free electron, namely $\vec{\vec{g}} = g_e\vec{\vec{1}}$. The so-called g tensor

shift, $\Delta\vec{\vec{g}} = \vec{\vec{g}} - g_e\vec{\vec{1}}$, thus vanishes, indicating that the influence of the other electrons in the molecule on the unpaired electron is neglected. At the lowest-order relativistic treatment, i.e., $\mathcal{O}(\alpha^2)$, which is usually considered in the derivation of the conventional formalism for nonrelativistic electronic g tensors, we obtain two corrections to the spin-Zeeman effect from first-order DGPT

$$\begin{aligned}\mathbf{H}_{\mathrm{RMC}} &= \langle \mathbf{0}|\hat{H}_{\mathrm{SZKE}}|\mathbf{0}\rangle \\ &= \mu_{\mathrm{B}} \sum_{k=x,y,z} B_k[\Sigma_{\mathrm{RMC}}]_{kk}\boldsymbol{\sigma}_k \\ &= \mu_{\mathrm{B}} \sum_{k=x,y,z} {}^{00}\Omega_{\mathrm{SZKE}} B_k \boldsymbol{\sigma}_k \end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{H}_{GC} &= \langle \mathbf{0}|\hat{H}_{\mathrm{GC(1e)}}|\mathbf{0}\rangle \\ &= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{GC}}]_{kl}\boldsymbol{\sigma}_l \\ &= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[{}^{00}\Omega_{\mathrm{GC(1e)}}]_{kl}\boldsymbol{\sigma}_l \end{aligned} \tag{12}$$

Taking into account the relationship between the electronic g tensor and the $\vec{\vec{\Sigma}}$ tensor (see eq 8), the terms described above in the $\vec{\vec{\Sigma}}$ expansion can readily be associated with the so-called mass velocity, $\Delta\vec{\vec{g}}_{\mathrm{RMC}} = 2\vec{\vec{\Sigma}}_{\mathrm{RMC}}$, and gauge, $\Delta\vec{\vec{g}}_{\mathrm{GC}} = 2\vec{\vec{\Sigma}}_{\mathrm{GC}}$, terms in the electronic g tensor shift. The remaining $\mathcal{O}(\alpha^2)$ corrections to the spin-Zeeman effect can be retrieved from second-order DGPT by combining the spin-Zeeman operator $\hat{H}_{\mathrm{SZ}}$ with the mass velocity $\hat{H}_{\mathrm{MV}}$, one-electron Darwin $\hat{H}_{\mathrm{DW}}$, and the one- and two-electron spin−orbit, $\hat{H}_{\mathrm{SO(1e)}}$ and $\hat{H}_{\mathrm{SO(2e)}}$, operators, which are all included in the perturbation $\hat{V}$ (see Table 1 for details). However, all these corrections vanish as $\hat{H}_{\mathrm{SZ}}$, being a pure spin operator, gives zero matrix elements between the electronic ground and excited states of the molecule. Therefore, at $\mathcal{O}(\alpha^2)$ only two corrections to the spin-Zeeman effect exist, namely the mass velocity and gauge corrections, if we assume that the perturbation $\hat{V}$ has the form defined in Table 1. Applying a similar procedure to the orbital Zeeman effect, we obtain one nonvanishing correction of $\mathcal{O}(\alpha^2)$ which originates from second-order DGPT, in which the orbital Zeeman effect operator $\hat{H}_{\mathrm{OZ}}$ is coupled to the spin−orbit interaction operators

$$\begin{aligned}\mathbf{H}_{\mathrm{OZ/SO}} &= \langle \mathbf{0}|\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SO}} + \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZ}}|\mathbf{0}\rangle \\ &= \frac{\mu_{\mathrm{B}}}{2} \sum_{k=x,y,z} B_k[\Lambda_{\mathrm{SO}}]_{kl}\boldsymbol{\sigma}_l \\ &= \frac{\mu_{\mathrm{B}}}{2} \sum_{k,l=x,y,z} B_k\left(4\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{OZ}}]_k[{}^{n0}\Omega_{\mathrm{SO}}]_l}{E_0 - E_n}\right)\boldsymbol{\sigma}_l \end{aligned} \tag{13}$$

where $\hat{H}_{\mathrm{SO}}$ includes both the one- and two-electron spin−orbit operators, $\hat{H}_{\mathrm{SO(1e)}}$ and $\hat{H}_{\mathrm{SO(2e)}}$, with the combined matrix element denoted as ${}^{nn'}\Omega_{\mathrm{SO}} = {}^{nn'}\Omega_{\mathrm{SO(1e)}} + {}^{nn'}\Omega_{\mathrm{SO(2e)}}$. We have in eq 13 also inserted the explicit form of the reduced resolvent operator $\hat{R}$ (see Appendix A) as well as matrix elements of the $\hat{H}_{\mathrm{OZ}}$ and $\hat{H}_{\mathrm{SO}}$ operators. Summing up all $\mathcal{O}(\alpha^2)$ corrections to the spin and orbital Zeeman effects, and extracting the corresponding contributions to the $\vec{\vec{\Lambda}}$ and $\vec{\vec{\Sigma}}$ tensors, we can according to eq 8 write the g tensor shifts as

$$\begin{aligned}\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2)) &= \Delta\vec{\vec{g}}_{\mathrm{SO}} + \Delta\vec{\vec{g}}_{\mathrm{RMC}} + \Delta\vec{\vec{g}}_{\mathrm{GC}} \\ &= \vec{\vec{\Lambda}}_{\mathrm{SO}} + 2\vec{\vec{\Sigma}}_{\mathrm{RMC}} + 2\vec{\vec{\Sigma}}_{\mathrm{GC}} \end{aligned} \tag{14}$$

which differs from the commonly used one only by having omitted the two-electron part of the gauge-correction term. We have here intentionally excluded the two-electron gauge operator from the perturbation operator $\hat{V}$ in our derivation, as the g tensor shift arising from this operator usually is considered to be negligible and therefore in most cases can be safely neglected (recently, an alternative view on the importance of the one- and two-electron gauge corrections to the g tensor shift has been discussed by Patchkovskii et al.[38]). Another reason for excluding this operator from our treatment stems from its two-electron nature, which makes the evaluation of the matrix elements a nontrivial task especially in the case of density functional theory. The same difficulties are also encountered in the computation of the $\Delta\vec{\vec{g}}_{\mathrm{SO}}$ term in the g tensor shift, which involves the two-electron spin−orbit operator $\hat{H}_{\mathrm{SO(2e)}}$. However, in this case the two-electron part of $\Delta\vec{\vec{g}}_{\mathrm{SO}}$ cannot be neglected, as this contribution along with its one-electron counterpart dominates the electronic g tensor shift of most molecules composed of main-group elements. One way to overcome this difficulty is to employ an effective one-electron spin−orbit operator in which the spin−orbit interaction screening effect of $\hat{H}_{\mathrm{SO(2e)}}$ is accounted for in an approximate manner. This approach would not only allow us to simplify the evaluation of the matrix elements but also to resolve the conceptional difficulty in the implementation of the $\Delta\vec{\vec{g}}_{\mathrm{SO}}$ term in density functional theory, as the formation of the two-particle density matrix, which is required for the computation of the two-electron part, can be avoided. In this paper, we will employ the atomic mean field approximation (AMFI)[39] for handling the two-electron spin−orbit operator matrix elements, as this method is well established and has previously been extensively used in computations of the electronic g tensor shifts[10,15,16,20,23−25] in the Kohn−Sham formalism. The choice of using the AMFI approximation is not only motivated by our desire to avoid problems associated with the construction of the two-particle density matrix in DFT but also dictated by current limitations in our quadratic response code, which cannot handle Breit−Pauli two-electron spin−orbit operators in the evaluation of several of the leading-order relativistic corrections to the electronic g tensor.

Having settled the treatment of the $\hat{H}_{\mathrm{SO(2e)}}$ operator in the calculation of $\Delta\vec{\vec{g}}_{\mathrm{SO}}$, let us now briefly review the computational cost of the various terms included in eq 14. An inspection of eqs 11−13 reveals that the most computationally expensive part of the $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$ evaluation is the determination of the spin−orbit contribution $\Delta\vec{\vec{g}}_{\mathrm{SO}}$, which requires the solution of a set of linear response equations. The remaining contributions to the g tensor shift are, on the other hand, straightforwardly evaluated as expectation values of the corresponding one-electron operators, and their computational cost is negligible compared to $\Delta\vec{\vec{g}}_{\mathrm{SO}}$. We note that spin-unrestricted and spin-restricted density functional linear response formalisms have been imple-

***Table 1.*** List of Breit−Pauli Hamiltonian Operators Included in Perturbation $\hat{V}$ and Their Matrix Elements in Degenerate Perturbation Theory

| label | order | operator description | $\langle \mathbf{n}\vert\hat{H}\vert\mathbf{n}'\rangle^{a,b,c}$ | ${}^{nn'}\Omega$ or $[{}^{nn'}\Omega]_k$ or $[{}^{nn'}\Omega]_{kl}{}^{c,d,e}$ |
|---|---|---|---|---|
| $\hat{H}_{SZ}$ | $\alpha^0$ | spin-Zeeman effect | $\mu_B^{nn'}\Omega_{SZ}\sum_{k=x,y,z} B_k\boldsymbol{\sigma}_k$ | $\delta_{nn'}$ |
| $\hat{H}_{OZ}$ | $\alpha^0$ | orbital Zeeman effect | $\mu_B\sum_{k=x,y,z}\mathbf{1}B_k[{}^{nn'}\Omega_{OZ}]_k$ | $\langle n\vert\sum_i [l_{iO}]_k\vert n'\rangle$ |
| $\hat{H}_{SZKE}$ | $\alpha^2$ | mass-velocity correction to spin-Zeeman effect | $\mu_B^{nn'}\Omega_{SZKE}\sum_{k=x,y,z} B_k\boldsymbol{\sigma}_k$ | $\alpha^2\langle n\vert\sum_i \nabla_i^2[s_i]_z\vert n'\rangle$ |
| $\hat{H}_{OZKE}$ | $\alpha^2$ | mass-velocity correction to orbital Zeeman effect | $\mu_B\sum_{k=x,y,z}\mathbf{1}B_k[{}^{nn'}\Omega_{OZKE}]_k$ | $\frac{\alpha^2}{2}\langle n\vert\sum_i [l_{iO}]_k\nabla_i^2\vert n'\rangle$ |
| $\hat{H}_{GC(1e)}$ | $\alpha^2$ | one-electron gauge correction to Zeeman effect | $\mu_B\sum_{k,l=x,y,z} B_k[{}^{nn'}\Omega_{GC(1e)}]_{kl}\boldsymbol{\sigma}_l$ | $\frac{\alpha^2}{2}\left\langle n\left\vert\sum_{i,N}\frac{\delta_{kl}(\vec{r}_{iN}\vec{r}_{iO})-[r_{iN}]_k[r_{iO}]_l}{r_{iN}^3}[s_i]_z\right\vert n'\right\rangle$ |
| $\hat{H}_{SO(1e)}$ | $\alpha^2$ | one-electron spin−orbit | $\sum_{k=x,y,z}[{}^{nn'}\Omega_{SO(1e)}]_k\boldsymbol{\sigma}_k$ | $\frac{\alpha^2}{2}\left\langle n\vert\sum_{i,N} Z_N\frac{[l_{iN}]_k}{r_{iN}^3}[s_i]_z\vert n'\right\rangle$ |
| $\hat{H}_{SO(2e)}$ | $\alpha^2$ | two-electron spin−orbit | $\sum_{k=x,y,z}[{}^{nn'}\Omega_{SO(2e)}]_k\boldsymbol{\sigma}_k$ | $-\frac{\alpha^2}{2}\left\langle n\left\vert\sum_{i\neq j}\frac{[l_{ij}]_k[s_i]_z+2[l_{ij}]_k[s_j]_z}{r_{ij}^3}\right\vert n'\right\rangle$ |
| $\hat{H}_{MV}$ | $\alpha^2$ | mass velocity | ${}^{nn'}\Omega_{MV}\mathbf{1}$ | $-\frac{\alpha^2}{8}\langle n\vert\sum_i \nabla_i^2\vert n'\rangle$ |
| $\hat{H}_{DW}$ | $\alpha^2$ | one-electron Darwin | ${}^{nn'}\Omega_{DV}\mathbf{1}$ | $\frac{\pi\alpha^2}{2}\langle n\vert\sum_{i,N} Z_N\delta(\vec{r}_{iN})\vert n'\rangle$ |

[a] Matrix element between two Kramer doublet states: $\vert n\rangle$ and $\vert n'\rangle$, which are constructed from degenerate doublet wave functions. [b] Here, we denoted Bohr magneton as $\mu_B = 1/2$. [c] Here, we use two notations for vector $\vec{A}$ components, namely $A_k$ or $[A]_k$, and one distinctive notation for tensor $\vec{\vec{T}}$ components $[T]_{kl}$. [d] Reduced matrix element, which is computed as high spin diagonal of $\langle n\vert\hat{H}\vert n'\rangle$ i.e., $\langle {}^2\mathrm{n};\ 1/2\vert\hat{H}\vert{}^2n';\ 1/2\rangle$. Detailed description of reduced matrix elements is given in Appendix B. [e] In definitions of reduced matrix elements, we employed following standard notations: $Z_N$ is the charge of nucleus $N$; $[s_i]_z$ is the z component of spin of electron $i$; $\vec{r}_{iN}$ is the position of electron $i$ with respect to nucleus $N$; $\vec{r}_{ij}$ is the position of electron $i$ with respect to electron $j$; $\vec{r}_{iO}$ is the position of electron $i$ with respect to gauge origin; $\vec{l}_{iO}$, $\vec{l}_{iN}$, and $\vec{l}_{ij}$ are the angular momentum of electron $i$ evaluated around the position of the gauge origin $O$, nucleus $N$, or electron $j$, respectively.

mented in various quantum chemistry programs in recent years,[10,12,14,35] and the evaluation of the nonrelativistic electronic g tensors, $\vec{\vec{g}}_{NR} = g_e\vec{\vec{1}} + \Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$, is therefore now becoming widely accessible. In this work, we will employ the spin-restricted density functional response formalism derived in ref 35 for evaluating the contributions to the electronic g tensor shift described above as well as higher-order terms (vide infra), as this approach is free from the spin contamination problem appearing in spin-unrestricted DFT approaches. This choice is motivated not only by the ability of the spin-restricted formalism to provide the pure spin states required by degenerate perturbation theory but also by the availability of a spin-restricted quadratic density functional response code,[36] which is needed for the evaluation of several of the $\mathcal{O}(\alpha^4)$ contributions to the electronic g tensor shift.

The leading-order relativistic corrections to the electronic g tensor is derived from degenerate perturbation theory by applying the same procedure as for the ordinary $\mathcal{O}(\alpha^2)$ contributions to $\Delta\vec{\vec{g}}$, although one in this case needs to go beyond the second-order DGPT in order to retrieve all relevant contributions to the $\vec{\vec{\Sigma}}$ and $\vec{\vec{\Lambda}}$ tensors. Since the perturbation $\hat{V}$ in this formalism only involves operators to

$\mathcal{O}(\alpha^2)$, the first corrections to the Zeeman effect of $\mathcal{O}(\alpha^4)$ will appear at second order of degenerate perturbation theory. Let us first consider the more complex case of the spin-Zeeman effect, to which we obtain two scalar relativistic corrections from second-order DGPT

$$\mathbf{H}_{\mathrm{RMC/SR}} = \langle \mathbf{0}|\hat{H}_{\mathrm{SZKE}}\hat{R}\hat{H}_{\mathrm{SR}} + \hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{SZKE}}|\mathbf{0}\rangle$$

$$= \mu_{\mathrm{B}} \sum_{k=x,y,z} B_k[\Sigma_{\mathrm{RMC/SR}}]_{kk}\boldsymbol{\sigma}_k$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k\left(2\sum_{n>0}\frac{{}^{0n}\Omega_{\mathrm{SZKE}}\,{}^{n0}\Omega_{\mathrm{SR}}}{E_0 - E_n}\right)\boldsymbol{\sigma}_k \quad (15)$$

$$\mathbf{H}_{\mathrm{GC/SR}} = \langle \mathbf{0}|\hat{H}_{\mathrm{GC(1e)}}\hat{R}\hat{H}_{\mathrm{SR}} + \hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{GC(1e)}}|\mathbf{0}\rangle$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{GC/SR}}]_{kk}\boldsymbol{\sigma}_k$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k\left(2\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{GC(1e)}}]_{kl}\,{}^{n0}\Omega_{\mathrm{SR}}}{E_0 - E_n}\right)\boldsymbol{\sigma}_l \quad (16)$$

which originate from the coupling of $\hat{H}_{\mathrm{SR}}$ with the operators defined by $\mathbf{H}_{\mathrm{RMC}}$ and $\mathbf{H}_{\mathrm{GC}}$. We have here introduced the combined $\hat{H}_{\mathrm{SR}}$ operator, which is responsible for the bulk part of the scalar relativistic corrections to $\hat{H}_0$, and includes the $\hat{H}_{\mathrm{MV}}$ and $\hat{H}_{\mathrm{DW}}$ operators. Its reduced matrix element is defined as ${}^{nn'}\Omega_{\mathrm{SR}} = {}^{nn'}\Omega_{\mathrm{MV}} + {}^{nn'}\Omega_{\mathrm{DW}}$. Apart from the scalar relativistic corrections to the spin-Zeeman effect described above, the second-order DGPT also gives rise to spin−orbit corrections

$$\mathbf{H}_{\mathrm{RMC/SO}} = \langle \mathbf{0}|\hat{H}_{\mathrm{SZKE}}\hat{R}\hat{H}_{\mathrm{SO}} + \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SZKE}}|\mathbf{0}\rangle$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{RMC/SO}}]_{kl}\boldsymbol{\sigma}_l$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k\left(2i\sum_{j=x,y,z}\epsilon_{kjl}\sum_{n>0}\frac{{}^{0n}\Omega_{\mathrm{SZKE}}[{}^{n0}\Omega_{\mathrm{SO}}]_j}{E_0 - E_n}\right)\boldsymbol{\sigma}_l \quad (17)$$

$$\mathbf{H}_{\mathrm{GC/SO}} = \langle \mathbf{0}|\hat{H}_{\mathrm{GC(1e)}}\hat{R}\hat{H}_{\mathrm{SO}} + \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{GC(1e)}}|\mathbf{0}\rangle$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{GC/SR}}]_{kl}\boldsymbol{\sigma}_l$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k \times \left(2i\sum_{j,j'=x,y,z}\epsilon_{jj'l}\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{GC(1e)}}]_{kj}[{}^{n0}\Omega_{\mathrm{SO}}]_{j'}}{E_0 - E_n}\right)\boldsymbol{\sigma}_l \quad (18)$$

in which instead of operators responsible for scalar relativistic effects, the spin−orbit interaction operator is coupled with $\hat{H}_{\mathrm{SZKE}}$ and $\hat{H}_{\mathrm{GC(1e)}}$. Here, we only consider doublet states in the derivation of the spin−orbit corrections to the spin-Zeeman effect and therefore neglect the contributions arising from quartet states in the DGPT expressions involving two or more electronic spin-dependent operators. This approximation is used throughout this paper and applied in the derivation of all contributions to the electronic g tensor shift which feature more than one spin-dependent operator. In addition to this approximation, we have also in the two latter equations, as well as other equations given below, employed the commutation and anticommutation relations of the Pauli matrices to reduce the equations to a final form featuring only a single Pauli matrix. From these four corrections to the spin-Zeeman effect, we can readily identify the second-order DGPT contributions to the $\vec{\vec{\Sigma}}$ tensor of $\mathcal{O}(\alpha^4)$, which can be interpreted as the spin−orbit, $\vec{\vec{\Sigma}}_{\mathrm{RMC/SO}}$ and $\vec{\vec{\Sigma}}_{\mathrm{GC/SO}}$, and scalar relativistic, $\vec{\vec{\Sigma}}_{\mathrm{RMC/SR}}$ and $\vec{\vec{\Sigma}}_{\mathrm{GC/SR}}$, corrections to the lower-order terms, $\vec{\vec{\Sigma}}_{\mathrm{RMC}}$ and $\vec{\vec{\Sigma}}_{\mathrm{GC}}$, in the $\vec{\vec{\Sigma}}$ tensor expansion.

In addition to these four corrections to the spin-Zeeman effect, two nonvanishing corrections are also obtained from third-order DGPT. One of these corrections is the so-called higher-order spin−orbit correction to the spin-Zeeman effect which only depends on the $\hat{H}_{\mathrm{SO}}$ and $\hat{H}_{\mathrm{SZ}}$ operators

$$\mathbf{H}_{\mathrm{SZ/SO}} = \langle \mathbf{0}|\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SZ}}\hat{R}\hat{H}_{\mathrm{SO}}|\mathbf{0}\rangle - \frac{1}{2}\{\langle \mathbf{0}|\hat{H}_{\mathrm{SZ}}|\mathbf{0}\rangle, \langle \mathbf{0}|\hat{H}_{\mathrm{SO}}\hat{R}^2\hat{H}_{\mathrm{SO}}|\mathbf{0}\rangle\}$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{SZ/SO}}]_{kl}\boldsymbol{\sigma}_l$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k\left(\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{SO}}]_k[{}^{n0}\Omega_{\mathrm{SO}}]_l}{(E_0 - E_n)^2} + \sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{SO}}]_l[{}^{n0}\Omega_{\mathrm{SO}}]_k}{(E_0 - E_n)^2} - 2\delta_{kl}\sum_{j=x,y,z}\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{SO}}]_j[{}^{n0}\Omega_{\mathrm{SO}}]_j}{(E_0 - E_n)^2}\right)\boldsymbol{\sigma}_l \quad (19)$$

In the above equation, as well as other similar equations, only the terms from third-order perturbation theory which give nonvanishing contributions to the final results are kept. The other contribution is of mixed "scalar relativistic/spin−orbit" nature and involves the $\hat{H}_{\mathrm{SR}}$ operator in addition to the $\hat{H}_{\mathrm{SO}}$ and $\hat{H}_{\mathrm{SZ}}$ operators

$$\mathbf{H}_{\mathrm{SZ/SO/SR}} = \langle \mathbf{0}|\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SZ}}\hat{R}\hat{H}_{\mathrm{SR}} + \hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{SZ}}\hat{R}\hat{H}_{\mathrm{SO}}|\mathbf{0}\rangle$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k[\Sigma_{\mathrm{SZ/SR}}]_{kl}\boldsymbol{\sigma}_l$$

$$= \mu_{\mathrm{B}} \sum_{k,l=x,y,z} B_k\left(2i\sum_{j=x,y,z}\epsilon_{kjl}\sum_{n>0}\frac{{}^{0n}\Omega_{\mathrm{SR}}[{}^{n0}\Omega_{\mathrm{SO}}]_j}{(E_0 - E_n)^2}\right)\boldsymbol{\sigma}_l \quad (20)$$

At $\mathcal{O}(\alpha^4)$ in degenerate perturbation theory, a total of six corrections to the spin-Zeeman effect with nonvanishing

contributions to the $\vec{\vec{\Sigma}}$ tensor is obtained, whereas in the work of Manninen et al.,[17] which employed conventional perturbation theory, only two corrections to the spin-Zeeman effect (see eqs 21−24 of ref 17), corresponding to the $\mathbf{H}_{\mathrm{RMC/SR}}$ and $\mathbf{H}_{\mathrm{GC/SR}}$ terms in our formalism, were obtained.

Let us now turn to the less complicated case of the orbital Zeeman effect, for which the application of DGPT leads to only three nonvanishing corrections of $\mathcal{O}(\alpha^4)$. One of these corrections arises from second-order DGPT and involves a kinetic energy correction to the orbital Zeeman effect operator $\hat{H}_{\mathrm{OZKE}}$ coupled to the spin−orbit operator in a similar fashion as in the $\mathbf{H}_{\mathrm{OZ/SO}}$ case

$$\begin{aligned}\mathbf{H}_{\mathrm{OZKE/SO}} &= \langle \mathbf{0}|\hat{H}_{\mathrm{OZKE}}\hat{R}\hat{H}_{\mathrm{SO}} + \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZKE}}|\mathbf{0}\rangle \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k[\Lambda_{\mathrm{SO/KE}}]_{kl}\boldsymbol{\sigma}_l \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k\left(4\sum_{n>0}\frac{[{}^{0n}\Omega_{\mathrm{OZKE}}]_k[{}^{n0}\Omega_{\mathrm{SO}}]_l}{(E_0-E_n)^2}\right)\boldsymbol{\sigma}_l\end{aligned} \tag{21}$$

The two remaining corrections are of a more complex nature and are obtained from third-order DGPT. One of these describes the scalar relativistic effects on the lower-order $\mathbf{H}_{\mathrm{OZ/SO}}$ correction and is defined as

$$\begin{aligned}\mathbf{H}_{\mathrm{OZ/SO/SR}} &= \langle \mathbf{0}|\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SR}} + \hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SO}} + \\ &\quad \hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZ}}|\mathbf{0}\rangle + \langle \mathbf{0}|\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{SO}} + \\ &\quad \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SR}}\hat{R}\hat{H}_{\mathrm{OZ}} + \hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SR}}|\mathbf{0}\rangle - \\ &\quad \frac{1}{2}\{\langle \mathbf{0}|\hat{H}_{\mathrm{SR}}|\mathbf{0}\rangle, \langle \mathbf{0}|\hat{H}_{\mathrm{SO}}\hat{R}^2\hat{H}_{\mathrm{OZ}} + \hat{H}_{\mathrm{OZ}}\hat{R}^2\hat{H}_{\mathrm{SO}}|\mathbf{0}\rangle\} \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k[\Lambda_{\mathrm{SO/SR}}]_{kl}\boldsymbol{\sigma}_l \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k \times \\ &\left(2\sum_{n,m>0}\frac{{}^{0n}\Omega_{\mathrm{SR}}([{}^{nm}\Omega_{\mathrm{OZ}}]_k[{}^{m0}\Omega_{\mathrm{SO}}]_l + [{}^{nm}\Omega_{\mathrm{OZ}}]_k[{}^{m0}\Omega_{\mathrm{SO}}]_l)}{(E_0-E_n)(E_0-E_m)} + \right. \\ &2\sum_{n,m>0}\frac{{}^{m0}\Omega_{\mathrm{SR}}([{}^{0n}\Omega_{\mathrm{OZ}}]_k[{}^{nm}\Omega_{\mathrm{SO}}]_l + [{}^{0n}\Omega_{\mathrm{SO}}]_k[{}^{nm}\Omega_{\mathrm{OZ}}]_l)}{(E_0-E_n)(E_0-E_m)} + \\ &2\sum_{n,m>0}\frac{{}^{nm}\Omega_{\mathrm{SR}}([{}^{0n}\Omega_{\mathrm{OZ}}]_k[{}^{m0}\Omega_{\mathrm{SO}}]_l + [{}^{0n}\Omega_{\mathrm{SO}}]_k[{}^{m0}\Omega_{\mathrm{OZ}}]_l)}{(E_0-E_n)(E_0-E_m)} - \\ &\left.4\sum_{n>0}\frac{{}^{00}\Omega_{\mathrm{SR}}([{}^{0n}\Omega_{\mathrm{OZ}}]_k[{}^{n0}\Omega_{\mathrm{SO}}]_l)}{(E_0-E_n)^2}\right)\boldsymbol{\sigma}_l\end{aligned} \tag{22}$$

whereas the last contribution describes the spin−orbit effect on the same lower-order correction $\mathbf{H}_{\mathrm{OZ/SO/SR}}$ and has the form

$$\begin{aligned}\mathbf{H}_{\mathrm{OZ/SO/SO}} &= \langle \mathbf{0}|\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SO}} + \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZ}}\hat{R}\hat{H}_{\mathrm{SO}} + \\ &\quad \hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{SO}}\hat{R}\hat{H}_{\mathrm{OZ}}|\mathbf{0}\rangle \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k[\Lambda_{\mathrm{SO/SO}}]_{kl}\boldsymbol{\sigma}_l \\ &= \frac{\mu_{\mathrm{B}}}{2}\sum_{k,l=x,y,z} B_k \times \\ &\left(2i\sum_{j,j'=x,y,z}\epsilon_{jj'l}\sum_{n,m>0}\frac{[{}^{0n}\Omega_{\mathrm{OZ}}]_k[{}^{nm}\Omega_{\mathrm{SO}}]_j[{}^{m0}\Omega_{\mathrm{SO}}]_{j'}}{(E_0-E_n)(E_0-E_m)} + \right. \\ &2i\sum_{j,j'=x,y,z}\epsilon_{jj'l}\sum_{n,m>0}\frac{[{}^{0n}\Omega_{\mathrm{SO}}]_j[{}^{nm}\Omega_{\mathrm{OZ}}]_k[{}^{m0}\Omega_{\mathrm{SO}}]_{j'}}{(E_0-E_n)(E_0-E_m)} + \\ &\left.2i\sum_{j,j'=x,y,z}\epsilon_{jj'l}\sum_{n,m>0}\frac{[{}^{0n}\Omega_{\mathrm{SO}}]_j[{}^{nm}\Omega_{\mathrm{SO}}]_{j'}[{}^{m0}\Omega_{\mathrm{OZ}}]_k}{(E_0-E_n)(E_0-E_m)}\right)\boldsymbol{\sigma}_l\end{aligned} \tag{23}$$

From the last three equations we can identify the corresponding contributions to the $\vec{\vec{\Lambda}}$ tensor of $\mathcal{O}(\alpha^4)$, which are all corrections to the lower-order contribution to $\vec{\vec{\Lambda}}_{\mathrm{SO}}$. More specifically, the spin−orbit correction is retrieved from $\mathbf{H}_{\mathrm{OZ/SO/SO}}$, whereas the active and passive scalar relativistic corrections are obtained from $\mathbf{H}_{\mathrm{OZ/KE/SO}}$ and $\mathbf{H}_{\mathrm{OZ/SO/SR}}$, respectively. The expression for the leading-order relativistic corrections to the g tensor shift can thus be summarized as

$$\begin{aligned}\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4)) &= \Delta\vec{\vec{g}}_{\mathrm{SO/KE}} + \Delta\vec{\vec{g}}_{\mathrm{SO/SR}} + \Delta\vec{\vec{g}}_{\mathrm{SO/SO}} + \Delta\vec{\vec{g}}_{\mathrm{RMC/SR}} + \\ &\quad \Delta\vec{\vec{g}}_{\mathrm{RMC/SO}} + \Delta\vec{\vec{g}}_{\mathrm{GC/SR}} + \Delta\vec{\vec{g}}_{\mathrm{GC/SO}} + \Delta\vec{\vec{g}}_{\mathrm{SZ/SR}} + \Delta\vec{\vec{g}}_{\mathrm{SZ/SO}} \\ &= \vec{\vec{\Lambda}}_{\mathrm{SO/KE}} + \vec{\vec{\Lambda}}_{\mathrm{SO/SR}} + \vec{\vec{\Lambda}}_{\mathrm{SO/SO}} + 2\vec{\vec{\Sigma}}_{\mathrm{RMC/SR}} + 2\vec{\vec{\Sigma}}_{\mathrm{RMC/SO}} + \\ &\quad 2\vec{\vec{\Sigma}}_{\mathrm{GC/SR}} + 2\vec{\vec{\Sigma}}_{\mathrm{RMC/SO}} + 2\vec{\vec{\Sigma}}_{\mathrm{SZ/SR}} + 2\vec{\vec{\Sigma}}_{\mathrm{SZ/SO}}\end{aligned} \tag{24}$$

where we have included all contributions to the $\vec{\vec{\Lambda}}$ and $\vec{\vec{\Sigma}}$ tensors of $\mathcal{O}(\alpha^4)$ according to eq 8. The relativistic g tensor can thus be defined as

$$\vec{\vec{g}} = \vec{\vec{g}}_{\mathrm{NR}} + \Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4)) = g_{\mathrm{e}}\vec{\vec{1}} + \Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2)) + \Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4)) \tag{25}$$

and includes the nonrelativistic g tensor corrected by $\Delta\vec{\vec{g}}$ ($\mathcal{O}(\alpha^4)$), which contains all the leading-order relativistic corrections. A detailed inspection of the various terms appearing in $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ indicates that the contributions to the g tensor shift can be separated into three clearly distinct types: (1) passive and active scalar relativistic corrections to $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$: $\Delta\vec{\vec{g}}_{\mathrm{SO/SR}}$, $\Delta\vec{\vec{g}}_{\mathrm{RMC/SR}}$, $\Delta\vec{\vec{g}}_{\mathrm{GC/SR}}$, and $\Delta\vec{\vec{g}}_{\mathrm{SO/KE}}$; (2) spin−orbit corrections to $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$: $\Delta\vec{\vec{g}}_{\mathrm{SO/SO}}$, $\Delta\vec{\vec{g}}_{\mathrm{RMC/SO}}$, and $\Delta\vec{\vec{g}}_{\mathrm{GC/SO}}$; (3) various corrections to the spin-Zeeman effect, involving the $\hat{H}_{\mathrm{SZ}}$ operator directly, namely $\Delta\vec{\vec{g}}_{\mathrm{SZ/SR}}$ and $\Delta\vec{\vec{g}}_{\mathrm{SZ/SO}}$.

The formulas in eq 24 for the evaluation of the relativistic g tensor contains the full range of relativistic effects originating from various one-electron operators in the Breit−Pauli Hamiltonian and provides a complete treatment of these effects in the g tensor shift to $\mathcal{O}(\alpha^4)$) for the selected form of perturbations $\hat{V}$ (see Table 1 for details). In the previous work by Manninen et al.,[17] only scalar relativistic contributions to the g tensor shift $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ were included—that is, the $\Delta\vec{\vec{g}}_{\mathrm{SO/SR}}$, $\Delta\vec{\vec{g}}_{\mathrm{RMC/SR}}$, $\Delta\vec{\vec{g}}_{\mathrm{GC/SR}}$, and $\Delta\vec{\vec{g}}_{\mathrm{SO/KE}}$ terms, and thus only a partial description of the

leading-order relativistic corrections to the electronic g tensor was considered in that work. Furthermore, in ref 17 all contributions to $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ which involved two or more electronic spin-dependent operators were also neglected. The formalism presented here has similarities to the "two-step" approach by Bolvin,[20] which includes the scalar relativistic corrections self-consistently and the spin−orbit corrections to the orbital and spin-Zeeman effect via a few-state model. More specifically, the $^2\vec{\vec{\Sigma}}$ and $^2\vec{\vec{\Lambda}}$ tensor in Bolvin's definition of the g tensor (see eq 26 of ref 20) can be identified as $\Delta\vec{\vec{g}}_{SZ/SO}$ and the sum of $\Delta\vec{\vec{g}}_{SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$, respectively. A detailed comparison of the contributions to the g tensor in the two approaches for the case of linear molecules are given in Appendix C. Since the "two-step" approach accounts for scalar relativistic effects self-consistently, this approach implicitly handles passive scalar relativistic contributions to the g tensor, which in our methodology is designated as $\Delta\vec{\vec{g}}_{SO/SR}$ and $\Delta\vec{\vec{g}}_{SZ/SR}$ in the expression for the g tensor shift. The methodology presented here can thus be considered a generalization, in different ways, of previous work[17,20] using perturbation theory to treat relativistic corrections to the g tensor. It provides a new and tractable formalism for handling leading-order scalar relativistic, spin−orbit and mixed corrections to the g tensor shifts.

Before discussing the results of our g tensor calculations for the $F_2^-$, $Cl_2^-$, $Br_2^-$, $I_2^-$ molecules including the relativistic corrections, we will comment on the usage of linear and quadratic response functions for the evaluation of various DGPT contributions to the electronic g tensor. A comparison of the DGPT expressions for evaluation of $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$ and $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$, and more specifically their orbital parts, with the spectral representation of linear and quadratic response functions, allows us to identify the corresponding response functions for computation of $\Delta\vec{\vec{g}}_{SO}$ as well as other contributions to the electronic g tensor, with the exception of the $\Delta\vec{\vec{g}}_{SO/SO}$ and $\Delta\vec{\vec{g}}_{SZ/SO}$ contributions, which cannot be defined by conventional response functions.

After now having established the applicability of response functions in DGPT calculations of electronic g tensors, we would also like to address one important practical point in the determination of linear and quadratic response functions involving triplet operators, which are an important part of calculations of the electronic g tensor. In linear and quadratic response functions of open-shell molecules, where only one of the perturbations is of triplet spin symmetry, the excited-state manifold is spanned by singlet excitations from the ground state. However, the perturbation operator itself retain its triplet nature and, consequently, its gradient, which enters the response equations, must be evaluated using the corresponding spin density, but not the total densities. Neglecting this fact is equivalent to reducing all triplet operators to singlet operators and leads, as we will show here, to an unphysical behavior. Unfortunately, because the generalizations required for triplet perturbations in open-shell molecules was implemented only recently for linear response functions,[6] and are applied here for the first time for quadratic response functions, the results of ref 17 suffer from this error. This lead in ref 17 to the conclusion that the scalar relativistic corrections are responsible for the large parallel g tensor shift in heavy dihalogen anion radicals, contrary to the findings of Malkin et al.,[18] as well as to Bolvin's results[20] obtained for other diatomic systems containing heavy elements, which both indicated that the spin−orbit interaction is responsible for this effect.

## III. Computational Details

Calculations of nonrelativistic g tensors of the dihalogen anion radicals, as well as the leading-order relativistic corrections to them, have been carried out using the methodology described in the previous section. The contributions to $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$ and $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ have been computed using spin-restricted density functional linear and quadratic response functions with the exception of the $\Delta\vec{\vec{g}}_{SO/SO}$ and $\Delta\vec{\vec{g}}_{SZ/SO}$ terms, for which we employed the sum-overstates (SOS) approach due to limitations in our density functional response code.[35,36] Apart from the DFT calculations, we have also carried out calculations at the spin-restricted open-shell Hartree−Fock (ROHF) level in order to demonstrate the importance of an appropriate treatment of triplet operators for the $\Delta\vec{\vec{g}}_{SO}$, $\Delta\vec{\vec{g}}_{RMC/SR}$, $\Delta\vec{\vec{g}}_{GC/SR}$, and $\Delta\vec{\vec{g}}_{SO/SR}$ contributions to the g tensor shift, which were investigated previously in ref 17 In all calculations we used the AMFI approximation for the evaluation of the $\hat{H}_{SO(2e)}$ operator matrix elements. For the matrix elements of the gauge-dependent operators, we employed a gauge origin centered at the electronic charge centroid of the molecule. For all calculations we used the geometries of the $X_2^-$ (X = F, Cl, Br, I) anion radicals from the work of Manninen et al.,[17] with the bond lengths in $F_2^-$, $Cl_2^-$, $Br_2^-$ and $I_2^-$ being 1.8916, 2.5570, 2.7943, and 3.1478 Å, respectively. These bond lengths were also used by Malkin et al.,[18] thus facilitating a direct comparison of our results with those of refs 17 and 18. Following the same reasoning, we employed a large uncontracted Huzinaga basis set (denoted here as HIVu, the details on this basis set can be found in ref 17), which was used to calculate the relativistic g tensor of the dihalogen anions by Manninen et al.[17] at the ROHF level. The HIVu basis set is furthermore sufficiently large to obtain near basis-set limit results for the dihalogen anion radicals.[17] In order to test the dependence of the g tensors and their relativistic corrections on the choice of exchange-correlation functional, we have carried out calculations using the local density Dirac−Vosko−Wilk−Nusair[40] (LDA), gradient-corrected Becke−Lee−Yang−Parr[41,42] (BLYP) and hybrid Becke3−Lee−Yang−Parr[40−44] (B3LYP) functionals. All calculations of linear and quadratic response functions as well as their residues were carried out using individually adjusted convergence thresholds, typically $10^{-8}$ for the response functions and $10^{-12}$ for the wave function, to ensure 0.01 ppm accuracy for each of the nonrelativistic as well as the leading-order relativistic contributions to electronic g tensor shift. All calculations were performed using a development version of the DALTON program.[45]

## IV. Results and Discussion

Dihalogen anion radicals are homonuclear diatomics with 15 valence electrons of which, in the $|X^2\Sigma_u^+\rangle$ ground state,[46,47]

***Table 2.*** Results of Nonrelativistic ($\mathcal{O}(\alpha^2)$) Calculations of Electronic g Tensor Shifts in Dihalogen Anion Radicals[a]

| | | $\Delta g_{\parallel}$ (in ppm) | | | $\Delta g_{\perp}$ (in ppm) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| molecule | method | RMC | GC | total | RMC | GC | SO(1e)[b] | SO(2e)[b] | total |
| $F_2^-$ | ROHF | −396 | 134 | −262 | −396 | 255 | 19558 | −6195 | 13222 |
| | LDA | −403 | 134 | −269 | −403 | 269 | 28605 | −8920 | 19551 |
| | BLYP | −402 | 134 | −268 | −402 | 269 | 26093 | −8128 | 17831 |
| | B3LYP | −402 | 134 | −269 | −402 | 266 | 24963 | −7804 | 17024 |
| | BP86[c] | −448 | 144 | −304 | −448 | 282 | 23949 | – | 23783 |
| | ROHF[d] | −425[e] | – | −425 | −454[e] | – | 15152[f] | – | 14698 |
| | MRCI[d] | −425[e] | – | −425 | −454[e] | – | 18406[f] | – | 17952 |
| | ROHF[g] | −558[h] | 165[h] | −393 | −558[h] | 316[h] | 24150 | – | 23908 |
| | MCSCF[g] | −493[h] | 150[h] | −343 | −493[h] | 293[h] | 22318 | – | 22118 |
| $Cl_2^-$ | ROHF | −286 | 154 | −132 | −286 | 272 | 31204 | −5644 | 25546 |
| | LDA | −305 | 155 | −150 | −305 | 302 | 52846 | −9360 | 43483 |
| | BLYP | −302 | 154 | −148 | −302 | 302 | 51164 | −9071 | 42092 |
| | B3LYP | −300 | 155 | −145 | −300 | 296 | 46864 | −8342 | 38518 |
| | BP86[c] | −321 | 156 | −165 | −321 | 305 | 48828 | – | 48812 |
| | ROHF[d] | −311[e] | – | −311 | −429[e] | – | 33255[f] | – | 32826 |
| | MRCI[d] | −311[e] | – | −331 | −429[e] | – | 41669[f] | – | 41240 |
| | ROHF[g] | −366[h] | 153[h] | −213 | −366[h] | 311[h] | 41524 | – | 41469 |
| | MCSCF[g] | −362[h] | 153[h] | −209 | −362[h] | 310[h] | 46004 | – | 45952 |
| $Br_2^-$ | ROHF | −310 | 266 | −44 | −310 | 451 | 112335 | −10690 | 101786 |
| | LDA | −331 | 264 | −67 | −331 | 501 | 202592 | −18894 | 183867 |
| | BLYP | −326 | 263 | −63 | −326 | 499 | 197890 | −18486 | 179576 |
| | B3LYP | −326 | 265 | −61 | −326 | 490 | 178645 | −16749 | 162060 |
| | BP86[c] | −326 | 258 | −68 | −326 | 506 | 188684 | – | 188864 |
| | ROHF[g] | −371[h] | 251[h] | −120 | −371[h] | 498[h] | 152050 | – | 152178 |
| | MCSCF[g] | −371[h] | 251[h] | −120 | −371[h] | 498[h] | 151615 | – | 151742 |
| $I_2^-$ | ROHF | −307 | 341 | 34 | −307 | 552 | 196220 | −12340 | 184125 |
| | LDA | −326 | 336 | 10 | −326 | 619 | 377954 | −23561 | 354687 |
| | BLYP | −319 | 333 | 14 | −319 | 616 | 369651 | −23096 | 346852 |
| | B3LYP | −320 | 337 | 17 | −320 | 604 | 330512 | −20689 | 310107 |
| | BP86[c] | −312 | 324 | 12 | −312 | 633 | 348526 | – | 348847 |
| | ROHF[g] | −348[h] | 306[h] | −42 | −348[h] | 604[h] | 282883 | – | 283139 |
| | MCSCF[g] | −348[h] | 306[h] | −42 | −348[h] | 604[h] | 302450 | – | 302706 |

[a] ROHF and DFT (LDA, BLYP and B3LYP functionals) calculations performed using Huz-IVu basis set. [b] The spin−orbit contribution to nonrelativistic g tensor i.e. $\Delta\vec{\vec{g}}_{SO}$ is splitted into one- and two-electron parts and $\hat{H}_{SO(2e)}$ matrix elements entering $\Delta\vec{\vec{g}}_{SO(2e)}$ computed using AMFI approximation. [c] Unrestricted DFT calculations from ref 17 with BP86 exchange-correlation functional. [d] ROHF and MCSCF calculations from refs 46 and 47. [e] Sum of first-order contributions, i.e., $\Delta g_{RMC} + \Delta g_{GC}$, where the last term includes both one- and two-electron parts. [f] Sum of second order spin−orbit contributions, i.e., $\Delta g_{SO(1e)} + \Delta g_{SO(2e)}$. [g] ROHF and MCSCF calculations from ref 17 [h] Computed using restricted-unrestricted approach, i.e., spin polarization is accounted for in these calculations.

14 electrons are distributed between the $\sigma_{g(u)}$ and $\pi_{g(u)}$ orbitals and the single unpaired electron is located in a $\sigma_u$ orbital (the typical valence electronic configuration is $\sigma_g^2\sigma_u^2\sigma_g^2\pi_u^4\pi_g^4\sigma_u^1$). The electronic structure and magnetic properties of the two lightest members of this family, $F_2^-$ and of $Cl_2^-$, have been extensively investigated by ab initio methods and the behavior of their electronic g tensors are well understood.[46,47] In contrast, the influence of relativity on the electronic g tensor of the heavier members of this anion family still remains an open question and two different views on the dominating mechanism of the relativistic corrections have been presented in the literature. On the basis of the results of MCSCF response theory calculations of $\mathcal{O}(\alpha^4)$ contributions to the electronic g tensors, Manninen et al.[17] advocated that the major relativistic corrections to the g tensors of $Br_2^-$ and $I_2^-$ are of scalar relativistic nature arising from the $\Delta\vec{\vec{g}}_{RMC/SR}$ and $\Delta\vec{\vec{g}}_{GC/SR}$ contributions to the g tensor shift. In contrast, a recent two-component DFT study by Malkin et al.[18] indicated that the relativistic effects to the electronic g tensors of $Br_2^-$ and $I_2^-$ are governed by higher-order spin−orbit corrections and that scalar relativistic effects are of minor importance. These disparate views on the influence of relativity on the electronic g tensors, and the inherent simplicity of the magnetic coupling responsible for the nonrelativistic g tensors, make the family of dihalogen anion compounds attractive as a test suite for validating the formalism developed in this paper for evaluating electronic g tensors.

**A. Nonrelativistic g Tensors of the $X_2^-$ (X = F, Cl, Br, I) Anion Radicals.** Before we go into an analysis of the relativistic effects on the electronic g tensors of the $X_2^-$ (X = F, Cl, Br, I) anion radicals, we first briefly discuss the nonrelativistic electronic g tensors and the interactions responsible for the parallel and perpendicular components of the g tensor shifts. The results of our DFT response calculations for the contributions of $\mathcal{O}(\alpha^2)$ to $\Delta\vec{\vec{g}}$, along with available results of previous ab initio and DFT calculations, are given in Table 2. As for other linear molecules, the parallel component of the electronic g tensor $g_{\parallel}$ of $X_2^-$ (X = F, Cl, Br, I), deviates only slightly from the free electron g factor, as the spin−orbit contribution is negligible (see eq 14). The $g_{\parallel}$ values are therefore entirely defined by the balance between two small contributions: the relativistic mass-velocity and gauge corrections to the spin-Zeeman effect, i.e. $[\Delta g_{RMC}]_{\parallel}$ and $[\Delta g_{GC}]_{\parallel}$. Overall, our DFT results agree very well with the BP86 results of Manninen et al.,[17] while our ROHF results gives systematically smaller $[\Delta g_{RMC}]_{\parallel}$ and $[\Delta g_{GC}]_{\parallel}$ (in terms of absolute values) compared

to their ROHF results. We rationalize this difference between these two virtually identical ROHF calculations by the fact that the restricted−unrestricted[48] formalism was used to compute these two contributions in ref 17. That is, in addition to the expectation values of the $\hat{H}_{RMC}$ and $\hat{H}_{GC(1e)}$ operators, the spin polarization contributions have been also computed. By comparing their and our results, we estimate that spin polarization mostly influences the relativistic mass-velocity contribution $[\Delta g_{RMC}]_{||}$, whereas the gauge correction $[\Delta g_{GC}]_{||}$ shows a significantly less pronounced dependence on spin polarization. Furthermore, the importance of spin polarization is reduced for the heavier members of the $X_2^-$ anion family (the estimated spin polarization contribution to $[\Delta g_{RMC}]_{||}$ is around 41% for $F_2^-$ and only around 13% for $I_2^-$, see Table 2 and ref 17). On the basis of these findings, we conclude that spin polarization can play a significant role for the nonrelativistic $g_{||}$ values of $F_2^-$ and $Cl_2^-$ (a detailed discussion of spin polarization effects in radicals with Σ-type ground states can be found in ref 49). Despite this, we choose to neglect spin polarization effects in the evaluation of the relativistic g tensors, as these effects cannot be accounted for systematically for all contributions included in the relativistic g tensor (see eq 25). Both our ROHF results as well as those of Manninen et al.[17] systematically underestimate the $\Delta g_{||}$ values compared to those obtained by Bruna and Grein[46,47] for $F_2^-$ and $Cl_2^-$, indicating that the two-electron part of $[\Delta g_{GC}]_{||}$, which have been accounted for in their work, is fairly large (in terms of absolute values) and almost cancel its one-electron counterpart for $F_2^-$ and $Cl_2^-$. In this respect, it is worth noting that the conventional view that the two-electron part of the gauge correction $\Delta\vec{\vec{g}}_{GC}$ is small and thus can be neglected does not hold for these dihalogen anion radicals. In this particular case, a more appropriate procedure would be to compute or neglect both contributions as suggested by Patchkovskii et al. in their recent work[38] on the evaluation of two-electron gauge corrections to electronic g tensors. However, in order to demonstrate the relativistic effects on $\Delta\vec{\vec{g}}_{GC}$, we kept its one-electron part in our calculations. These results underscore the importance of a balanced description of first order (in the perturbation theory sense) contributions to $\Delta g_{||}$ in linear molecules. To summarize, the $g_{||}$ components of the $X_2^-$ (X = F, Cl, Br, I) compounds deviate at $\mathcal{O}(\alpha^2)$ only slightly from the free-electron g factor and is defined by the subtle balance between mass-velocity and gauge corrections to the spin-Zeeman effect. A comparative analysis of our results and those of Bruna and Grein corroborates the findings by Patchkovskii et al.[38] that the widely accepted view about the magnitude of the gauge correction to the spin-Zeeman effect cannot be considered trustworthy, and that the two-electron part of this correction actually can be a very important contribution to nonrelativistic $g_{||}$ of linear molecules. Therefore, in cases when the second-order spin−orbit contribution vanishes or is comparable to the first-order contributions, both the one- and two-electron parts of the gauge correction to the spin-Zeeman effect should be considered, and one cannot a priori assess the relative importance of these parts.

In contrast to $g_{||}$, the perpendicular component of the electronic g tensor $g_\perp$ differs significantly from the free-electron g factor for all dihalogen anion radicals, and this deviation increases dramatically when going from $F_2^-$ to $I_2^-$ (see Table 2). The behavior of $g_\perp$ for the dihalogen anion radicals is governed by the dominating second-order spin−orbit contribution $\Delta\vec{\vec{g}}_{SO}$, and the first-order contributions, which determine $g_{||}$, are negligible for this component of the nonrelativistic g tensor. However, we note that the first-order contributions show a similar trend as for $g_{||}$, with the exception of the $Br_2^-$ and $I_2^-$ anion radicals, for which the gauge contribution $[\Delta g_{GC}]_\perp$ becomes larger than the mass-velocity contribution $[\Delta g_{RMC}]_\perp$. For the dominant second-order one- and two-electron spin−orbit contributions, $[\Delta g_{SO(1e)}]_\perp$ and $[\Delta g_{SO(2e)}]_\perp$, our ROHF and DFT values agree well with previous MRCI and BP86 results (see Table 2), demonstrating the good performance of spin-restricted density functional response theory. We also note that there is only a moderate dependence of $[\Delta g_{SO}]_\perp$ on the exchange-correlation functionals used in the calculations. It is worth noting that the magnetic coupling, which gives rise to the spin−orbit contributions to $g_\perp$ in the $X_2^-$ radicals, is particularly simple—the major contribution originates from the $|1^2\Pi_u\rangle$ ($\pi_u \leftarrow \sigma_u$) excited state. Therefore, an accurate description of $g_\perp$ can be achieved by considering a single excited state for the electronic g tensor shift (a detailed discussion of this topic is presented in Appendix C, where a two-state model for evaluating the electronic g tensor in linear molecules is derived). According to Bruna and Grein,[46,47] the ratio between the angular momentum matrix element $\langle X^2\Sigma_u^+|\hat{H}_{OZ}|1^2\Pi_u\rangle$ and the $|X^2\Sigma_u^+\rangle \rightarrow |1^2\Pi_u\rangle$ excitation energy remains almost constant going from $F_2^-$ to $Cl_2^-$ and, consequently, the increase of the spin−orbit matrix element $\langle X^2\Sigma_u^+|\hat{H}_{SO}|1^2\Pi_u\rangle$ determines the changes in $\Delta g_\perp$. According to our calculations, the same trend holds relatively well also for $Br_2^-$ and $I_2^-$ and, consequently, the ratio between $\Delta g_\perp$ for two members of the dihalogen anion family is approximately equal to the ratio between their spin−orbit matrix elements $\langle X^2\Sigma_u^+|\hat{H}_{SO}|1^2\Pi_u\rangle$, although this correspondence between ratios is most accurate for neighboring $X_2^-$ radical pairs, such as $F_2^-$ and $Cl_2^-$, or $Br_2^-$ and $I_2^-$. The perpendicular component of the nonrelativistic electronic g tensor of the $X_2^-$ (X = F, Cl, Br, I) radicals can thus be rationalized by a two-state model as suggested by Bruna and Grein.[46,47] As the spin−orbit interaction plays a major role for the perpendicular component of the g tensor shift, it is interesting to look at the importance of the one- and two-electron spin−orbit contributions. A quick inspection of Table 2 reveals that the two-electron spin−orbit contribution to $\Delta g_\perp$ is only important for $F_2^-$ (around 31% of the $[\Delta g_{SO(1e)}]_\perp$ contribution) and $Cl_2^-$ (around 18% of the $[\Delta g_{SO(1e)}]_\perp$ contribution), whereas the importance of this contribution decreases for the heavier members of this anion family. In fact, $[\Delta g_{SO(2e)}]_\perp$ amounts only to about 10% and 6% of the $[\Delta g_{SO(1e)}]_\perp$ contribution (in absolute value) for $Br_2^-$ and $I_2^-$, respectively. In agreement with previous work on the dihalogen compounds,[10] we find that the neglect of the two-electron spin−orbit contribution to $g_\perp$ for the two

***Table 3.*** Leading-Order Relativistic Contributions to the g Tensor Shift of Dihalogen Anion Radicals[a]

| | | $\Delta g_{\parallel}$ (in ppm) | | | | | $\Delta g_{\perp}$ (in ppm) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| molecule | method | RMC/SR | GC/SR | SZ/SO | SO/SO | total | RMC/SR | GC/SR | SO/SR | SO/KE | SZ/SO | total |
| $F_2^-$ | ROHF | −0.29 | 0.02 | −151.68 | 75.84 | −76.11 | −0.29 | 0.42 | 60.56 | −3.87 | −75.84 | −19.02 |
| | LDA | −0.27 | 0.01 | −297.52 | 148.76 | −149.02 | −0.27 | 0.40 | 88.18 | −4.63 | −149.00 | −65.32 |
| | BLYP | −0.27 | 0.01 | −260.90 | 130.45 | −137.71 | −0.27 | 0.39 | 79.96 | −4.06 | −130.58 | −54.56 |
| | B3LYP | −0.28 | 0.02 | −238.16 | 119.08 | −119.34 | −0.28 | 0.40 | 76.50 | −4.14 | −119.20 | −46.44 |
| | ROHF[b] | −143.74 | 25.12 | – | – | −118.62 | −143.74 | 25.00 | 97.59 | −13.22 | – | −34.37 |
| $Cl_2^-$ | ROHF | −0.81 | −0.01 | −778.10 | 389.05 | −389.87 | −0.81 | 2.52 | 580.88 | −3.98 | −389.77 | 188.84 |
| | LDA | −0.83 | −0.01 | −1915.82 | 957.91 | −958.75 | −0.83 | 2.50 | 947.70 | −8.24 | −958.93 | −17.80 |
| | BLYP | −0.84 | −0.01 | −1790.36 | 895.18 | −896.03 | −0.84 | 2.45 | 909.30 | −7.41 | −896.14 | 7.36 |
| | B3LYP | −0.84 | −0.01 | −1533.14 | 766.57 | −767.42 | −0.84 | 2.49 | 843.74 | −6.93 | −767.39 | 71.07 |
| | ROHF[b] | −2140.66 | 374.72 | – | | −1765.94 | −2140.66 | 374.12 | 814.78 | −104.12 | – | −1055.88 |
| $Br_2^-$ | ROHF | −4.95 | −0.14 | −14528.38 | 7264.19 | −7269.28 | −4.95 | 25.91 | 13605.86 | −6.72 | −7271.93 | 6348.17 |
| | LDA | −4.22 | −0.34 | −37932.00 | 18966.00 | −18970.56 | −4.22 | 25.20 | 23362.54 | −43.58 | −18986.20 | 4353.74 |
| | BLYP | −4.16 | −0.32 | −35848.66 | 17924.33 | −17928.81 | −4.16 | 24.66 | 22639.70 | −37.96 | −17943.42 | 4678.82 |
| | B3LYP | −4.44 | −0.26 | −30265.44 | 15132.72 | −15137.42 | −4.44 | 25.24 | 20711.50 | −33.44 | −15148.84 | 5550.02 |
| | ROHF[b] | −44733.01 | 7759.88 | – | – | −36973.13 | −44733.01 | 7803.51 | 17532.21 | −2587.48 | – | −21984.77 |
| $I_2^-$ | ROHF | −12.52 | −1.29 | −63648.60 | 31824.30 | −31838.11 | −12.52 | 92.51 | 67229.06 | 52.63 | −31858.20 | 35503.48 |
| | LDA | −9.25 | −2.11 | −172355.92 | 86177.96 | −86189.32 | −9.25 | 90.18 | 118863.22 | −77.49 | −86269.74 | 32596.92 |
| | BLYP | −9.12 | −2.07 | −160874.78 | 80437.39 | −80448.58 | −9.12 | 87.94 | 115020.68 | −57.81 | −80523.06 | 34518.63 |
| | B3LYP | −10.11 | −1.86 | −135532.26 | 67766.13 | −67778.10 | −10.11 | 90.17 | 104979.68 | −40.53 | −67838.30 | 37180.91 |
| | ROHF[b] | −253913.11 | 43944.08 | – | – | −209969.03 | −253913.11 | 43919.74 | 83520.87 | −15425.12 | – | −141897.62 |

[a] ROHF and DFT (LDA, BLYP, and B3LYP functionals) calculations performed using Huz-IVu basis set. [b] ROHF calculation results from ref 17.

heaviest members of the $X_2^-$ compounds is well justified, whereas this is not the case for the two lightest members.

To summarize, the nonrelativistic g tensor of the dihalogen anion radicals behaves similar to the g tensor of other linear molecules consisting of main group elements—that is, the parallel component of the g tensor is close to the free-electron g factor and the perpendicular component deviates significantly from the free-electron g factor, and this deviation depends almost exclusively on the magnitude of the spin−orbit coupling between the $|X^2\Sigma_u^+\rangle$ ground state and the $|1^2\Pi_u\rangle$ excited state.

**B. Leading-Order Relativistic Corrections to the g Tensors of $X_2^-$ (X = F, Cl, Br, I) Anion Radicals.** The calculated leading-order relativistic corrections to the electronic g tensors of the dihalogen anion radicals are presented in Table 3. We start the analysis of the results by first considering the scalar relativistic contributions to $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$, which have already been considered by Manninen et al.[17] A quick inspection of the $\Delta\vec{\vec{g}}_{\mathrm{RMC/SR}}$ and $\Delta\vec{\vec{g}}_{\mathrm{GC/SR}}$ contributions reveals that these passive scalar relativistic corrections give negligible contributions to the parallel and perpendicular components of the g tensor shift for the entire dihalogen anion family, and even for the heaviest member of these compounds $I_2^-$, it only constitutes about 0.1−2% of the corresponding nonrelativistic contributions, $\Delta\vec{\vec{g}}_{\mathrm{RMC}}$ and $\Delta\vec{\vec{g}}_{\mathrm{GC}}$, from which these corrections originate (see eqs 15 and 16). A similar situation is also encountered in the case of the active scalar correction to $[\Delta g_{\mathrm{SO}}]_{\perp}$, namely $[\Delta g_{\mathrm{SO/KE}}]_{\perp}$, which is numerically insignificant and does not exceed 0.1% of $[\Delta g_{\mathrm{SO}}]_{\perp}$ for $I_2^-$. In contrast, $[\Delta g_{\mathrm{SO/SR}}]_{\perp}$ is nonnegligible and increases from about 0.06 to 118.9 ppt going from $F_2^-$ to $I_2^-$, depending somewhat on the choice of exchange-correlation functional. For $Br_2^-$ and $I_2^-$, this correction accounts for an increase of about 13% and 35% of the g tensor shift compared to the nonrelativistic $\Delta g_{\perp}$, respectively. Consequently, only $\Delta\vec{\vec{g}}_{\mathrm{SO/SR}}$ gives a sizable contribution to $\Delta g_{\perp}$ and is thus responsible for the enhancement of this component in $Br_2^-$ and $I_2^-$, whereas in the case of $\Delta g_{\parallel}$, all corrections are negligible. We thus conclude that for compounds consisting of main group elements, scalar relativistic corrections to the nonrelativistic g tensor, given by the $\Delta\vec{\vec{g}}_{\mathrm{RMC/SR}}$, $\Delta\vec{\vec{g}}_{\mathrm{GC/SR}}$ and $\Delta\vec{\vec{g}}_{\mathrm{SO/KE}}$ terms in eq 24, can safely be neglected, as contributions arising from these corrections are in general small. For radicals containing heavy elements, on the other hand, passive scalar relativistic corrections to the spin−orbit term in the nonrelativistic g tensor, namely $\Delta\vec{\vec{g}}_{\mathrm{SO/SR}}$, is significant and must be accounted for in calculations of the electronic g tensor of such species.

These findings on the importance of the leading-order scalar relativistic corrections to the electronic g tensors of the dihalogen anion radicals are in marked contrast to the results of ref 17 and is due to the incorrect ROHF and MCSCF response code used in that work, as also noted in the previous section in the case of the second-order spin−orbit contribution to the nonrelativistic g tensor shift. In order to highlight the impact of this difference in calculations of the relativistic corrections, we have tabulated our recomputed ROHF values of the $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ contributions together with the results of ref 17 in Table 3. A direct comparison of these results reveals significant differences. For example, for the case of $Cl_2^-$, the scalar relativistic correction to $[\Delta g_{\mathrm{RMC}}]_{\perp}$ is in ref 17 found to be about −2.14 ppt, i.e., roughly 6 times larger than $[\Delta g_{\mathrm{RMC}}]_{\perp}$ itself, whereas our calculations predict the scalar relativistic correction to $[\Delta g_{\mathrm{RMC}}]_{\perp}$ to be −0.08 ppt, i.e., about 0.2% of $[\Delta g_{\mathrm{RMC}}]_{\perp}$. We will therefore not consider the results of Manninen et al.[17] in the remaining discussion.

Let us now turn to the next type of relativistic corrections to the electronic g tensor shift, namely the various spin−orbit corrections, which were omitted in ref 17 but which have been considered in part by Bolvin in connection with the "two-step" formalism.[20] The spin−orbit corrections to the mass-velocity and gauge terms in $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^2))$, $\Delta\vec{\vec{g}}_{\mathrm{RMC/SO}}$ and $\Delta\vec{\vec{g}}_{\mathrm{GC/SO}}$, vanish for the dihalogen anion radicals and are therefore not tabulated in Table 3. We point out that

$\Delta\vec{\vec{g}}_{RMC/SO}$ is a traceless antisymmetric tensor (see eqs 24 and 17) and does not contribute to the total isotropic g tensor shift of any molecule, not only in the specific case of $X_2^-$ (X = F, Cl, Br, I). The $\Delta\vec{\vec{g}}_{GC/SO}$, on the other hand, is in general expected to give nonnegligible contributions of $\mathcal{O}(\alpha^4)$ to $\Delta\vec{g}$, whereas in the case of dihalogen anion radicals it is numerically insignificant due to the structure of the excited states in these systems. In addition to these corrections, the so-called higher-order spin−orbit correction to the orbital Zeeman effect, $\Delta\vec{\vec{g}}_{SO/SO}$, also appears to $\mathcal{O}(\alpha^4)$. This correction gives a positive contribution to $\Delta g_{||}$ (see Table 3), but vanishes for $\Delta g_{\perp}$. The $\Delta\vec{\vec{g}}_{SO/SO}$ can thus be interpreted as the higher-order counterpart of $\Delta\vec{\vec{g}}_{SO}$, which requires two spin−orbit operators to mediate the interaction of the electrons with the external magnetic field via the $\hat{H}_{OZ}$ operator. The data in Table 3 shows that $[\Delta g_{SO/SO}]_{||}$ increases from 0.1 to 86.2 ppt going from the lightest to the heaviest members of the dihalogen anion radicals family, and already for $Cl_2^-$ it is several times bigger than the nonrelativistic $\Delta g_{||}$. However, since $[\Delta g_{SO/SO}]_{||}$ is positive, it cannot be responsible for the large negative $\Delta g_{||}$ observed experimentally in the $Br_2^-$ and $I_2^-$ radicals and it therefore only has a screening effect—in fact, it effectively moderates other large negative relativistic corrections to $\Delta g_{||}$ (vide infra).

In order to identify these large negative contributions, let us turn to the last class of relativistic corrections to the g tensor included in the DGPT approach, the corrections that depend explicitly on the spin-Zeeman operator (the last two terms in eq 24). The first of these contributions, $\Delta\vec{\vec{g}}_{SZ/SR}$, is similar to $\Delta\vec{\vec{g}}_{RMC/SO}$ in the sense of being an antisymmetric traceless tensor that does not contribute to the total g tensor shift. $\Delta\vec{\vec{g}}_{SZ/SO}$, on the other hand, gives rise to sizable contributions to both components of $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$ (see Table 3). For linear molecules with a Σ-type ground state, this correction is distributed in a 2:1 ratio between $\Delta g_{||}$ and $\Delta g_{\perp}$ as required by molecular symmetry (see Appendix C for details). The $[\Delta g_{SZ/SO}]_{||}$ is of moderate size for $F_2^-$ and $Cl_2^-$, but for the heaviest members of the dihalogen anion radical family, it increases significantly, being −18.97 and −86.18 ppt for $Br_2^-$ and $I_2^-$, respectively. $\Delta\vec{\vec{g}}_{SZ/SO}$ is therefore one of the largest (in terms of absolute values) relativistic corrections to the electronic g tensor shift and is responsible for the large negative parallel component of the g tensor shift of $Br_2^-$ and $I_2^-$. This result is in agreement with the findings of Malkin et al.,[18] which indicated that the spin−orbit effects determine $\Delta g_{||}$ in these radicals, since $\Delta\vec{\vec{g}}_{SZ/SO}$ originates from the spin−orbit interaction operator and is interpreted as a higher-order spin−orbit correction to the spin-Zeeman effect (see eq 19).

Let us now examine the collective effect of all these corrections to $\Delta\vec{\vec{g}}$ for the dihalogen anion radicals. For the parallel component of the g tensor, two relativistic corrections, $[\Delta g_{SZ/SO}]_{||}$ and $[\Delta g_{SO/SO}]_{||}$, are significant and they obey the relationship $[\Delta g_{SZ/SO}]_{||} \approx -1/2[\Delta g_{SO/SO}]_{||}$. The total relativistic correction to $\Delta g_{||}$ is therefore negative and approximately equal to $1/2[\Delta g_{SZ/SO}]_{||}$, and it is dominated almost exclusively by higher-order spin−orbit interaction contributions to the spin and orbital Zeeman effects. The collective relativistic correction to $\Delta g_{||}$ becomes nonnegligible already for $Cl_2^-$, and it increases steeply going to the heavier members of the dihalogen anion radical family. In this respect, it is worth to note that the nonrelativistic formalism for evaluating electronic g tensors fails already for $\Delta g_{||}$ of $Cl_2^-$, and an account of higher-order spin−orbit corrections to the Zeeman effect becomes essential for a correct prediction of $\Delta g_{||}$. The higher-order spin−orbit corrections thus become competitive in magnitude compared to the nonrelativistic contributions to the g tensor shift even in molecules containing relatively light elements if the lower-order spin−orbit correction $\Delta\vec{\vec{g}}_{SO}$ vanishes. For the perpendicular component of the g tensor shift of the dihalogen anion radicals, we encounter a different situation as both scalar-relativistic and spin−orbit corrections give sizable contributions to $\Delta g_{\perp}$ in this case. However, $[\Delta g_{SO/SR}]_{\perp}$ and $[\Delta g_{SZ/SO}]_{\perp}$ are of comparable magnitude but bear the opposite sign, and the total relativistic correction to $\Delta g_{\perp}$ is consequently relatively small, reflecting the fact that $[\Delta g_{SO/SR}]_{\perp}$ increases more rapidly compared to $[\Delta g_{SZ/SO}]_{\perp}$ when going from $F_2^-$ to $I_2^-$. Thus, the leading-order relativistic correction to the perpendicular component of the g tensor shift is of moderate size and, for example in the case of $I_2^-$, is responsible for an increase of $\Delta g_{\perp}$ by roughly 10−11% compared to the nonrelativistic $\Delta g_{\perp}$ value. In contrast to $\Delta g_{||}$, the relativistic correction to $\Delta g_{\perp}$ must therefore be accounted for in molecules consisting of the heaviest elements in the periodic table.

Before concluding the discussion of the leading-order relativistic corrections to the electronic g tensors, we would like to consider in more detail a computational aspect of the evaluation of $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$. These contributions are computed using a sum-overstates approach, including up to 15 excited states of $|^2\Pi_u\rangle$ type in the perturbation theory summation of the $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$ corrections. As expected, for all dihalogen anion radicals, a single excited state dominates the contribution to $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$, and this contribution accounts for more than 99% of the total value of the computed corrections for all $X_2^-$ (X = F, Cl, Br, I) compounds. Thus, our SOS expansion for $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$ corrections are well converged and including more excited states in the SOS expansion is not expected to change these values further. We would like to stress, however, that the good convergence of the SOS expansion is specific to the dihalogen anion radicals and is not necessarily transferable to other molecular systems, and thus one should carefully check the convergence of the SOS expansion when evaluating the $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$ corrections in order to avoid introducing significant errors in the determination of the electronic g tensor shifts. We would also like to point out that the approximate relation between $[\Delta g_{SZ/SO}]_{||}$ and $[\Delta g_{SO/SO}]_{||}$, i.e., $[\Delta g_{SZ/SO}]_{||} \approx -1/2[\Delta g_{SO/SO}]_{||}$, only holds for dihalogen anion radicals, as the reduced matrix elements of the orbital angular momentum between the $|^2\Pi_u\rangle$ states for these molecules obeys $[^{nm}\Omega_{OZ}]_z$ ($n \neq m$) vanish and $|[^{mm}\Omega_{OZ}]_z| \approx 1$, indicating that the $|^2\Pi_u\rangle$ states are dominated by contributions from pure $p_x$ and $p_y$ atomic orbitals. Considering the complications described above in the computation of the $\Delta\vec{\vec{g}}_{SZ/SO}$ and $\Delta\vec{\vec{g}}_{SO/SO}$ corrections to electronic g tensor shifts, it is evident that the determination

***Table 4.*** Nonrelativistic and Relativistic g Tensor Shifts of $X_2^-$ (X = F, Cl, Br, I) and Their Dependence on the Exchange-Correlation Functional[a]

| | | $\Delta g_{\|}$ (in ppt) | | | $\Delta g_{\perp}$ (in ppt) | | | | | $\Delta g_{\|}$ (in ppt) | | | $\Delta g_{\perp}$ (in ppt) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| molecule | method | nonrel[a] | rel[b] | total | nonrel[a] | rel[b] | total | molecule | method | nonrel[a] | rel[b] | total | nonrel[a] | rel[b] | total |
| $F_2^-$ | ROHF | −0.26 | −0.08 | −0.34 | 13.22 | −0.02 | 13.20 | $Cl_2^-$ | ROHF | −0.13 | −0.39 | −0.52 | 25.55 | 0.19 | 24.74 |
| | LDA | −0.27 | −0.15 | −0.42 | 19.55 | −0.07 | 19.48 | | LDA | −0.15 | −0.96 | −1.11 | 43.48 | −0.02 | 43.46 |
| | BLYP | −0.27 | −0.13 | −0.40 | 17.83 | −0.06 | 17.77 | | BLYP | −0.15 | −0.90 | −1.05 | 42.09 | 0.01 | 42.10 |
| | B3LYP | −0.27 | −0.12 | −0.39 | 17.02 | −0.05 | 16.97 | | B3LYP | −0.15 | −0.77 | −0.92 | 38.52 | 0.07 | 35.59 |
| | BP86[c] | −0.30 | – | −0.30 | 23.78 | – | 23.78 | | BP86[c] | −0.17 | – | −0.17 | 48.81 | – | 48.81 |
| | ROHF[d] | −0.43 | – | −0.43 | 14.70 | – | 14.70 | | ROHF[d] | −0.31 | – | −0.31 | 32.83 | – | 32.83 |
| | MRCI[d] | −0.43 | – | −0.43 | 17.95 | – | 17.95 | | MRCI[d] | −0.31 | – | −0.31 | 41.24 | – | 41.24 |
| | BP86/DK-1e[e] | 0.18[f] | – | 0.18 | 17.18[f] | – | 17.18 | | BP86/DK-1e[e] | 0.18[f] | – | 0.18 | 39.88[f] | – | 39.88 |
| | BP86/DK-2e[e] | −0.62[f] | – | −0.62 | 16.28[f] | – | 16.28 | | BP86/DK-2e[e] | −0.92[f] | – | −0.92 | 38.38[f] | – | 38.38 |
| | expt[g] | – | – | −0.52 | – | – | 16.18 | | expt[g] | – | – | −1.72 | – | – | 13.81 |
| | expt[h] | – | – | −0.32 | – | – | 19.08 | | expt[h] | – | – | −0.82 | – | – | 42.18 |
| $Br_2^-$ | ROHF | −0.31 | −7.27 | −7.58 | 101.79 | 6.35 | 108.14 | $I_2^-$ | ROHF | 0.03 | −31.84 | −31.83 | 184.13 | 35.50 | 219.63 |
| | LDA | −0.33 | −18.97 | −19.30 | 183.39 | 4.35 | 187.74 | | LDA | 0.01 | −86.19 | −86.18 | 354.69 | 32.60 | 387.29 |
| | BLYP | −0.33 | −17.93 | −18.26 | 179.58 | 4.68 | 184.26 | | BLYP | 0.01 | −80.44 | −80.43 | 346.85 | 34.52 | 381.37 |
| | B3LYP | −0.33 | −15.14 | −15.47 | 162.06 | 5.55 | 167.61 | | B3LYP | 0.02 | −67.77 | −67.75 | 310.11 | 37.18 | 347.29 |
| | BP86[c] | −0.01 | – | −0.01 | 188.86 | – | 188.86 | | BP86[c] | 0.01 | – | 0.01 | 348.85 | – | 348.85 |
| | BP86/DK-1e[e] | 0.18[f] | – | 0.18 | 178.58[f] | – | 178.58 | | BP86/DK-1e[e] | 0.28[f] | – | 0.28 | 369.98[f] | – | 369.98 |
| | BP86/DK-2e[e] | −17.82[f] | – | −17.82 | 151.68[f] | – | 151.68 | | BP86/DK-2e[e] | −75.62[f] | – | −75.62 | 243.68[f] | – | 243.68 |
| | expt[h] | – | – | −23.22 | – | – | 171.78 | | expt[h] | – | – | −143.92 | – | – | 307.68 |

[a] Nonrelativistic g tensor shifts evaluated according to eq 14. Computation carried out at ROHF and at DFT level (LDA, BLYP, and B3LYP functionals) in Huz-IVu basis set. [b] Relativistic g tensor shifts evaluated according to eqs 14 and 24. Computation carried out at ROHF and at DFT level (LDA, BLYP, and B3LYP functionals) in Huz-IVu basis set. [c] Unrestricted DFT calculations from ref 17 with BP86 exchange-correlation functional. [d] ROHF and MCSCF calculations from refs46 and 47. [e] Douglas−Kroll one- and two-component calculation results from ref 18. The one-component results denoted as DK-1e and two-component results denoted as DK-2e, respectively. [f] Total contribution to electronic g tensor shift, includes relativistic effects. [g] Experimental data for $F_2^-$ and $Cl_2^-$ obtained in neon[50] and argon[51] matrices, respectively. [h] Experimental data from NaX (X = F, Cl, Br, I) host crystal.[51]

of these correction is the most difficult part of calculating $\Delta\vec{\vec{g}}(\mathcal{O}(\alpha^4))$, as it requires a detailed knowledge of the excited states domain structure and a careful consideration of the sum-overstates expansion convergence.

To summarize, of the $\mathcal{O}(\alpha^4)$ contributions to $\Delta\vec{\vec{g}}$ defined in eq 24, only the scalar relativistic and spin−orbit corrections $\Delta\vec{\vec{g}}_{SO/SR}$, $\Delta\vec{\vec{g}}_{SO/SO}$, and $\Delta\vec{\vec{g}}_{SZ/SO}$, are important and need to be included in the calculation of electronic g tensors of compounds containing Br or I atoms, or other heavy elements. Another important aspect to note in the evaluation of the relativistic corrections to $\Delta\vec{\vec{g}}$ is the large differences between the corrections obtained in the ROHF and DFT calculations. These differences are likely to be due to the triplet instability problem that is often encountered in Hartree−Fock calculations involving triplet perturbations. The ROHF method can therefore not be recommended for calculations of relativistic corrections to electronic g tensors, and if one wants to pursue the ab initio track, MCSCF response theory will probably be required.

In Table 4, we have collected the nonrelativistic and relativistic g tensor shifts of the dihalogen radicals together with the previous one-component and two-component Douglas−Kroll DFT results of Malkin et al.,[18] as well as available experimental data. The most striking effect of the relativistic corrections to $\Delta\vec{\vec{g}}$ is observed for the parallel component of the g tensor shift, which increases by several orders of magnitude compared to its nonrelativistic value for the heaviest members of the dihalogen anion radicals family considered here. The relativistic effects are less pronounced on the perpendicular component of the g tensor shift, and amounts only to an increase by 10−11% compared to its nonrelativistic value for $I_2^-$. Our DFT results are in good agreement with the two-component Douglas-Kroll DFT results of Malkin et al.[18] (denoted as BP86/DK-2e in Table 4) for both components of the electronic g tensor shift. This result is as expected, as both methods account for the scalar relativistic and spin−orbit corrections to the nonrelativistic g tensors. The only difference between our results and those of Malkin et al.[18] is observed for $\Delta g_{\perp}$ of $Br_2^-$ and $I_2^-$, where our DGPT approach produces systematically larger $\Delta g_{\perp}$ values compared to the BP86/DK-2e method. This discrepancy is most likely caused by a different balance between the scalar-relativistic and spin−orbit corrections featured in these two approaches. Another interesting point to note is that the one-component Douglas−Kroll DFT method of Malkin et al.[18] (denoted as BP86/DK-1e in Table 4), which includes only leading-order scalar relativistic corrections, fails to correctly predict the increase of $\Delta g_{\|}$ going from $F_2^-$ to $I_2^-$. This result is in agreement with our findings that the scalar relativistic effects are negligible for the parallel component of the g tensor shift of the dihalogen anion radicals and that the results of Maninnen et al.,[17] which contradicts this statement, is caused by computational artifacts. Finally, we would like to point out that a comparison of the BP86/DK-1e and BP86/DK-2e results obtained in ref 18 indicates that the inclusion of higher-order spin−orbit corrections effectively reduces the magnitude of $\Delta g_{\perp}$ on $Br_2^-$ and $I_2^-$, and the same effect is also observed in our calculations (see the $[\Delta g_{SO/SR}]_{\perp}$ and $[\Delta g_{SZ/SO}]_{\perp}$ corrections in Table 3). Our DGPT approach therefore predicts a behavior of the g tensor shifts of the dihalogen anion radicals in agreement with the two-component Douglas−Kroll DFT approach, and both these very different methodologies give a consistent picture of the physical mechanisms responsible for the dominant relativistic corrections to the electronic g tensor of these compounds.

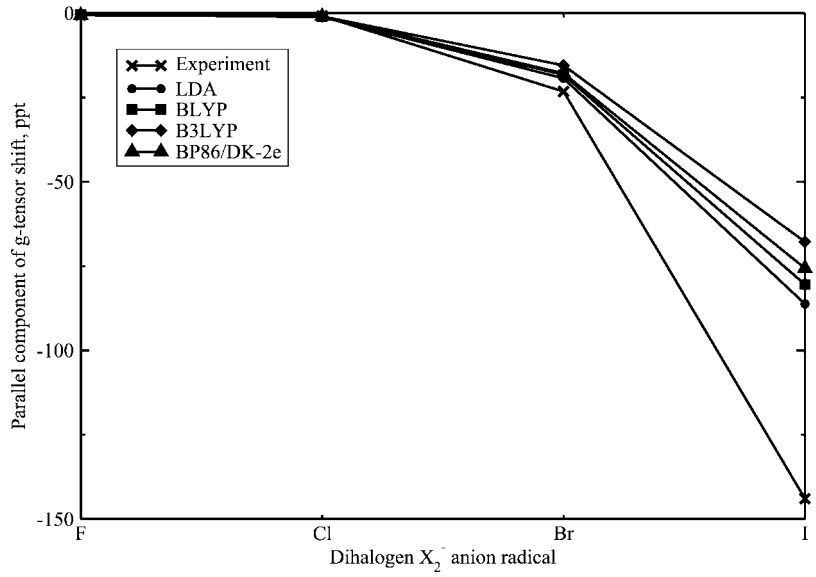



***Figure 1.*** Parallel component of the electronic g tensor shift of dihalogen anion radicals.

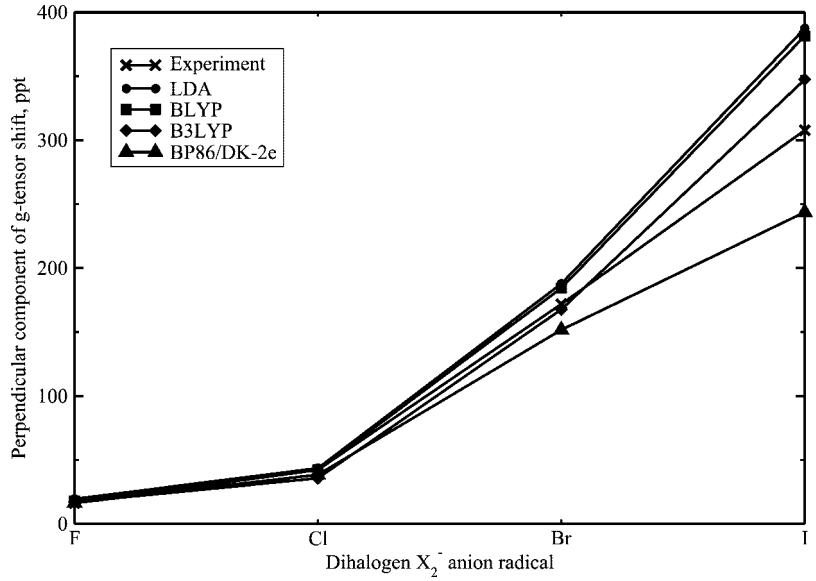



***Figure 2.*** Perpendicular component of the electronic g tensor shift of dihalogen anion radicals.

Let us now turn to an assessment of the performance of the DGPT approach with respect to available experimental data. In Figure 1, we have plotted the parallel component of the g tensor shift for all dihalogen anion radicals obtained with the LDA, BLYP, and B3LYP functionals along with the BP86/DK-2e results of ref 18 as well as experimental data. From this plot it is evident that the DGPT approach reproduces the experimentally observed trend in $\Delta g_{\parallel}$ for the whole series of dihalogen anion radicals independently of the choice of exchange-correlation functional. The best agreement with the experimental results is obtained using the LDA functional, as it consistently gives the smallest $|X^2\Sigma_u^+\rangle \rightarrow |1^2\Pi_u\rangle$ excitation energies and in this way facilitates the enhancement of the $[\Delta g_{SZ/SO}]_{\parallel}$ and $[\Delta g_{SO/SO}]_{\parallel}$ corrections (see Table 3). The BLYP and B3LYP functionals systematically predict smaller (in terms of absolute values) $\Delta g_{\parallel}$ values for all $X_2^-$ (X = F, Cl, Br, I) compounds. The BLYP results are in close agreement with the BP86/DK-2e results of Malkin et al.[18] The accuracy of the DGPT approach in predicting $\Delta g_{\parallel}$ is satisfactory for the first three members of the dihalogen anion radical family, but we underestimate $\Delta g_{\parallel}$ of $I_2^-$ approximately by a factor of 2 (in terms of absolute values), as can be seen from Figure 1. This discrepancy between theory and experiment is most likely caused by the complete neglect of environmental effects in our calculations, which play an important role in electronic g-tensor shifts as have been observed for the lightest members of the dihalogen anions family (see Bruna and Grein's MRCI and ROHF results in refs 46 and 47).

The perpendicular components of the g tensor shift of the $X_2^-$ (X = F, Cl, Br, I) compounds are plotted in Figure 2. As for $\Delta g_{\parallel}$, $\Delta g_{\perp}$ obtained with the DGPT approach qualitatively reproduce the experimental data for the whole series of dihalogen anion radicals, independently of the choice of exchange-correlation functional employed in the calculations. Of the different exchange-correlation functionals, B3LYP has the closest agreement with experiment as well as to the results of ref 18. LDA in this case systematically overesti-

mates $\Delta g_\perp$, whereas BLYP gives results inbetween LDA and B3LYP. The difference in performance of the exchange-correlation functionals for the parallel and perpendicular components of the g tensor shift does not allow us to select a functional most suited for evaluating the relativistic g tensor using the DGPT approach. However, based on the good performance of the B3LYP functional for $\Delta g_\perp$ in $X_2^-$ (X = F, Cl, Br, I), we prefer this exchange-correlation functional over BLYP or LDA, as it provides more consistent results for the electronic g tensor shifts and excitation energies in the dihalogen anion radicals and is expected to give results of similar quality also for other compounds containing Br or I atoms. To summarize, the DGPT approach presented in this paper in combination with spin-restricted open-shell density functional theory is capable of predicting the leading-order relativistic corrections to the electronic g tensors of dihalogen anion radicals as well as providing a clear picture of the physical mechanisms responsible for the observed relativistic effects. The DGPT approach is thus an attractive alternative approach to the two-component approach developed by Malkin et al.[18] for the calculation of relativistic g tensors in molecules containing heavy elements.

## V. Conclusions

The current work is a step in the development of perturbation theory based approaches for the calculation of the leading-order relativistic corrections to electronic g tensors. The approach developed here is based on degenerate perturbation theory and is formulated for molecules with an orbitally nondegenerate doublet ground state. The approach accounts for all relativistic corrections of $\mathcal{O}(\alpha^4)$ to the electronic g tensor, arising from the one-electron part of the Breit−Pauli Hamiltonian—that is, the scalar relativistic and spin−orbit corrections to the spin- and orbital-Zeeman effects. The formalism developed is in its present form limited to molecules with a doublet ground state, but it can relatively easily be extended to handle systems with ground states of different multiplicity by substituting the conventional Pauli matrices with generalized Pauli matrices, and rederiving the final expression for each term in eq 24. The DGPT approach can be implemented at the ab initio or density functional level of theory, provided that the selected method is capable of producing pure spin states. In this work, we have implemented this approach at the spin-restricted open-shell density functional theory level, as our main aim is to investigate large molecular systems containing heavy elements, which are beyond the reach of conventional ab initio methods such as MCSCF. In addition to the full DGPT formalism, we have also developed a simplified two-state model for the approximate calculation of electronic g tensors of linear molecules including leading-order relativistic corrections. We believe that both approaches will become useful tools for the investigation and analysis of EPR spectra of compounds containing heavy elements, where an account of relativistic effects is of critical importance.

The application of the DGPT approach to the dihalogen anion radicals shows that the leading-order relativistic corrections allows us to explain the variation of both g tensor shift components in these compounds, and we achieve satisfactory agreement between theory and available experimental data. The results obtained indicate that the spin−orbit interaction is responsible for the large negative parallel component of the g tensor shift in $Br_2^-$ and $I_2^-$ and that both scalar relativistic and spin−orbit effects are important for the perpendicular components of these two radicals. These findings are in a good agreement with the results of Malkin et al.[18] and also demonstrate that the conclusions of Manninen et al.[17] are incorrect due to an incorrect handling of triplet operators in their calculations. The controversy about the dominant physical mechanism responsible for the g tensor shift of the dihalogen anion radicals has thus been resolved.

**Acknowledgment.** This work was supported by a grant from the Swedish Infrastructure Committee (SNIC) for the project "Multiphysics Modeling of Molecular Materials", SNIC 023/07-18. This work has received support from the Norwegian Research Council through a Centre of Excellence Grant (Grant No. 179568/V30), a YFF grant to K.R. (Grant No. 162746/V00), as well as through a grant of computer time from the Norwegian Supercomputing Program. This work has also received support from NordForsk (Grant No. 070253).

## Appendix A: Degenerate Perturbation Theory for Computation of Molecular Properties

In open-shell molecular systems described by a nonrelativistic Hamiltonian, the ground state along with the excited states are degenerate with respect to the total electronic spin degrees of freedom, and consequently the evaluation of electronic spin-dependent properties, such as electronic g tensors, requires the application of a perturbation theory which can handle degeneracies of this kind. One such approach is the degenerate perturbation theory developed by Löwdin in the middle of the previous century.[52] In this approach, the degeneracy of the unperturbed states is tackled by employing the reduced resolvent operator technique, which effectively splits states of the molecular system into sets of nondegenerate state domains and ensures that the perturbation theory summations are free from divergences caused by the degeneracies. The original formulation of DGPT by Löwdin focused on the evaluation of various corrections to the electronic energy of a degenerate ground state, and little attention has been given to the treatment of molecular properties in this approach. In order to extend this approach, we have, in the spirit of response theory,[53] reformulated the DGPT equations in a form more convenient for the direct computation of arbitrary molecular properties in the presence of a perturbation $\hat{V}$.

Collecting all degenerate substates of the ground-state into a row matrix $|\mathbf{0}\rangle = (|^{2S+1}0;-m\rangle|^{2S+1}0;-m+1\rangle...|^{2S+1}0;m\rangle)$, and defining the conjugated state as a column matrix $\langle\mathbf{0}| = (\langle^{2S+1}0;-m|\langle^{2S+1}0;-m+1|...\langle^{2S+1}0;m|)^T$, we can in analogy with ordinary response theory expand a molecular property given by an operator $\hat{A}$ as

$$\mathbf{A} = \mathbf{A}^{(0)} + \mathbf{A}^{(1)} + \mathbf{A}^{(2)} + ... \qquad \text{(A.1)}$$

where we instead of the conventional response functions have introduced response matrices which span the degenerate

substates manifold. From the Löwdin expressions for evaluating first-, second-, and third-order corrections to the energy of the degenerate ground state, we can write down the following expressions for determining the zeroth-, first-, and second-order response matrices

$$\mathbf{A}^{(0)} = \langle \mathbf{0}|\hat{A}|\mathbf{0}\rangle$$
$$\mathbf{A}^{(1)} = \langle \mathbf{0}|\hat{A}\hat{R}\hat{V} + \hat{V}\hat{R}\hat{A}|\mathbf{0}\rangle$$
$$\mathbf{A}^{(2)} = \langle \mathbf{0}|\hat{A}\hat{R}\hat{V}\hat{R}\hat{V}|\mathbf{0}\rangle - \frac{1}{2}\{\mathbf{V}^{(0)}, \langle \hat{\mathbf{0}}|\hat{A}\hat{R}^2\hat{V}|\mathbf{0}\rangle\} +$$
$$\langle \mathbf{0}|\hat{V}\hat{R}\hat{A}\hat{R}\hat{V}|\mathbf{0}\rangle - \frac{1}{2}\{\mathbf{A}^{(0)}, \langle \hat{\mathbf{0}}|\hat{V}\hat{R}^2\hat{V}|\mathbf{0}\rangle\} + \langle \mathbf{0}|\hat{V}\hat{R}\hat{V}\hat{R}\hat{A}|\mathbf{0}\rangle -$$
$$\frac{1}{2}\{\mathbf{V}^{(0)}, \langle \mathbf{0}|\hat{V}\hat{R}^2\hat{A}|\mathbf{0}\rangle\} \quad \text{(A.2)}$$

where we have introduced the reduced resolvent operator $\hat{R}$, which projects out the ground-state manifold from the perturbation theory summations and is defined in terms of the unperturbed Hamiltonian $\hat{H}_0$ of the system as

$$\hat{R} = \frac{\hat{1} - |\mathbf{0}\rangle\langle\mathbf{0}|}{E_0 - \hat{H}_0} \quad \text{(A.3)}$$

These expressions for evaluating the property **A** are in general applicable to open-shell molecular systems with a ground state of arbitrary multiplicity and allow us to treat the degeneracy with respect to spin degrees of freedom in a uniform way at different orders in the perturbation theory expansion. Furthermore, for electronic spin-independent operators $\hat{A}$ and $\hat{V}$, the response matrices reduce to a diagonal form and, for each $|^{2S+1}0;m\rangle$ component of the degenerate ground state, the response matrix elements become identical to the ordinary response functions for evaluating static molecular properties.

## Appendix B: Reduced Matrix Elements for Doublet States

In this Appendix we present the details underlying the different reduced matrix elements of the spin rank one operators, such as the spin−orbit operator, appearing in Table 1. The matrix element between two Kramer's doublet states, $|\mathbf{n}\rangle$ and $|\mathbf{n}'\rangle$, for these operators can be expanded in terms of reduced matrix elements ${}^{nn'}\Omega_k$ and the Pauli matrices which, together with the two-dimensional unit matrix, form a complete basis in the space of two-dimensional square matrices

$$\langle \mathbf{0}|\hat{H}|\mathbf{0}\rangle = \sum_{k=x,y,z} {}^{nn'}\Omega_k \boldsymbol{\sigma}_k \quad \text{(B.1)}$$

The reduced matrix elements ${}^{nn'}\Omega_k$ entering the above equation can be determined by comparing the corresponding nonvanishing matrix elements on the right- and left-hand sides, i.e.

$$\langle {}^2n;m|\hat{H}_k|{}^2n';m'\rangle = {}^{nn'}\Omega_k[\boldsymbol{\sigma}_k]_{mm'} \quad \text{(B.2)}$$

and inserting the explicit expression for the $\langle {}^2n;m|\hat{H}_k|{}^2n';m'\rangle$ and $[\sigma_k]_{mm'}$ matrix elements. In order to accomplish this task, we transform all quantities in the latter equality into spherical coordinates, and further reduce the $\hat{H}$ matrix elements by applying the Wigner−Eckart theorem in spin space

$$\langle {}^2n;m|\hat{H}|{}^2n';m'\rangle = \sum_{\mu=-1,0,1} (-1)^{\mu}\langle {}^2n||\hat{H}_{-\mu}(\vec{s})||{}^2n'\rangle C^{(1/2)1(1/2)}_{m'\mu m} \quad \text{(B.3)}$$

where $C^{j_1j_2j}_{m_1m_2m}$ are Clebsch−Gordan coefficients and $\langle {}^2n||\hat{H}_{-\mu}(\vec{s})||{}^2n'\rangle$ are the reduced matrix elements which are independent of the spin projection quantum numbers. The operator in this expression refers in spin space to the operator as a whole, and the remaining index $\mu$ refers only to the spatial part (e.g., the orbital angular momentum part of the spin−orbit operator). A detailed discussion of the evaluation of matrix elements of this kind can be found in ref 54. In order to compute the matrix elements $\langle {}^2n||\hat{H}_{-\mu}(\vec{s})||{}^2n'\rangle$, we have for convenience selected the high-spin reference states components, $|{}^2n;{}^1/_2\rangle$ and $|{}^2n';{}^1/_2\rangle$, and this choice leads to the following expression for this matrix element

$$\langle {}^2n||\hat{H}_{-\mu}(\vec{s})||{}^2n'\rangle = \langle {}^2n;\frac{1}{2}|\hat{H}_{-\mu}(s_0)|{}^2n';\frac{1}{2}\rangle / C^{(1/2)1(1/2)}_{(1/2)0(1/2)} \quad \text{(B.4)}$$

which in turn allows us to obtain the final expression for evaluating the operator $\hat{H}$ matrix elements in spherical coordinates

$$\langle {}^2n;m|\hat{H}_{\mu}|{}^2n';m'\rangle = \sum_{\mu=-1,0,1} (-1)^{\mu}\langle {}^2n;\frac{1}{2}|\hat{H}_{-\mu}(s_0)|{}^2n';\frac{1}{2}\rangle C^{(1/2)1(1/2)}_{m'\mu m} / C^{(1/2)1(1/2)}_{(1/2)0(1/2)} \quad \text{(B.5)}$$

Taking this result into account, we can rewrite eq B.2, which defines the reduced matrix element ${}^{nn'}\Omega_k$ in Cartesian coordinates system, into an equivalent equation in spherical coordinates

$$\langle {}^2n;\frac{1}{2}|\hat{H}_{-\mu}(s_0)|{}^2n';\frac{1}{2}\rangle C^{(1/2)1(1/2)}_{m'\mu m} / C^{(1/2)1(1/2)}_{(1/2)0(1/2)} = {}^{nn'}\Omega_{-\mu}[\boldsymbol{\sigma}_{\mu}]_{mm'} \quad \text{(B.6)}$$

Since the values of $[\boldsymbol{\sigma}_{\mu}]_{mm'}$ can be easily obtained from the relationship

$$\frac{1}{2}[\boldsymbol{\sigma}_{\mu}]_{mm'} = C^{(1/2)1(1/2)}_{m'\mu m} / 2C^{(1/2)1(1/2)}_{(1/2)0(1/2)} \quad \text{(B.7)}$$

we can after some algebraic manipulations write the reduced matrix elements used in the $\langle n|\hat{H}|n'\rangle$ expansion as

$${}^{nn'}\Omega_{\mu} = \langle {}^2n;\frac{1}{2}|\hat{H}_{\mu}(s_0)|{}^2n';\frac{1}{2}\rangle \quad \text{(B.8)}$$

in spherical coordinates, or as

$${}^{nn'}\Omega_k = \langle {}^2n;\frac{1}{2}|\hat{H}_k(s_z)|{}^2n';\frac{1}{2}\rangle \quad \text{(B.9)}$$

in Cartesian coordinates. These reduced matrix elements can be found in Table 1.

## Appendix C: Two-State Model for Electronic g Tensors of Linear Molecules

The behavior of the electronic g tensors in linear molecules with a doublet ground state of $|{}^2\Sigma\rangle$ type is frequently

***Table 5.*** Comparison of g Tensor Shift of Dihalogen Anion Radicals (in ppt) Obtained Using Full DGPT and "Two-State" Approaches[a]

| | | two-state model | | DGPT | | | | two-state model | | DGPT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| molecule | functional | $\Delta g_{\|\|}$ | $\Delta g_{\perp}$ | $\Delta g_{\|\|}$ | $\Delta g_{\perp}$ | molecule | functional | $\Delta g_{\|\|}$ | $\Delta g_{\perp}$ | $\Delta g_{\|\|}$ | $\Delta g_{\perp}$ |
| $F_2^-$ | LDA | −0.15 | 20.51 | −0.42 | 19.48 | $Cl_2^-$ | LDA | −0.96 | 47.38 | −1.11 | 43.46 |
| | BLYP | −0.13 | 18.75 | −0.40 | 17.77 | | BLYP | −0.88 | 47.52 | −1.05 | 42.10 |
| | B3LYP | −0.12 | 17.91 | −0.39 | 16.97 | | B3LYP | −0.75 | 42.93 | −0.92 | 35.59 |
| $Br_2^-$ | LDA | −18.76 | 188.43 | −19.30 | 187.74 | $I_2^-$ | LDA | −85.56 | 339.76 | −86.18 | 387.29 |
| | BLYP | −17.62 | 181.97 | −18.26 | 184.26 | | BLYP | −79.57 | 329.51 | −80.43 | 381.37 |
| | B3LYP | −14.97 | 168.48 | −15.47 | 167.61 | | B3LYP | −67.20 | 307.63 | −67.75 | 347.29 |

[a] DFT (LDA, BLYP, and B3LYP functionals) calculations performed using Huz-IVu basis set.

governed by a single excited $|^2\Pi\rangle$ type state, which gives the dominant contribution to both the perpendicular and parallel components of the electronic g tensor shift. This feature of the electronic g tensor of linear molecules has been observed many times in theoretical investigations of diatomics consisting of main group elements,[5,11,15,47] and recently similar trends have been verified also for diatomics containing heavy elements.[20] In view of these findings, a simple two-state model for an approximate estimation of the electronic g tensor shift in linear molecules is desirable, as it not only allows us to predict the magnitude of the electronic g tensor shift but also provides valuable insight into the physical mechanisms responsible for this shift.

We will start the development of the two-state model by assuming that the bulk part of the parallel and perpendicular components of the g tensor is defined by the contribution from a single $|^2\Pi\rangle$ type state, which is one of the low-lying excited states in linear molecules. Similarly to Bolvin,[20] we will represent the orbitally doubly degenerate $\{^2\Pi$ state as two complex components $|^2\Pi_+\rangle$ and $|^2\Pi_-\rangle$. Under these conditions, we will apply degenerate perturbation theory for evaluating the relativistic electronic g tensor shift, taking into account only the $|^2\Sigma\rangle$ ground state and one $|^2\Pi\rangle$ excited state. As prescribed by the DGPT formalism, we first form a Kramers pair for the ground state $|\mathbf{\Sigma}\rangle = (|^2\Sigma;^1/_2\rangle|^2\Sigma;-^1/_2\rangle)$ and two row vectors for the $|^2\Pi\rangle$ state, namely $|\mathbf{\Pi}_+\rangle = (|^2\Pi_+;^1/_2\rangle|^2\Pi_+;-^1/_2\rangle)$ and $|\mathbf{\Pi}_-\rangle = (|^2\Pi_-;^1/_2\rangle|^2\Pi_-;-^1/_2\rangle)$. Taking into account only passive scalar relativistic and spin−orbit corrections to the orbital and spin-Zeeman interactions, i.e., limiting the perturbation operator to $\hat{H}_{SZ}$, $\hat{H}_{OZ}$, $\hat{H}_{SR} = \hat{H}_{MV} + \hat{H}_{DW}$ and $\hat{H}_{SO} = \hat{H}_{SO(1e)} + \hat{H}_{SO(2e)}$, we will obtain nonvanishing contributions to the electronic g tensor shift (see eqs 14 and 24) from $\Delta\vec{\vec{g}}_{SO}$, $\Delta\vec{\vec{g}}_{SO/SR}$, $\Delta\vec{\vec{g}}_{SO/SO}$, and $\Delta\vec{\vec{g}}_{SZ/SO}$. Before proceeding with the actual computation of these contributions to the g tensor shift, let us introduce the following definitions of the orbital Zeeman effect and energy-weighted spin−orbit interaction matrix elements between the $|^2\Sigma\rangle$ and $|^2\Pi\rangle$ states:

$$\langle\mathbf{\Sigma}|\hat{H}_{OZ}|^2\mathbf{\Pi}_{\pm}\rangle = \mu_B \sum_{k=x,y,z} B_k[^{\Sigma\Pi_{\pm}}\Omega_{OZ}]_k = \mu_B(B_xL \pm iB_yL)\mathbf{1} \tag{C.1}$$

$$\langle\mathbf{\Pi}_{\pm}|\hat{R}\hat{H}_{SO}|\mathbf{\Sigma}\rangle = \sum_{k=x,y,z} \frac{[^{\Pi_{\pm}\Sigma}\Omega_{SO}]_k\boldsymbol{\sigma}_k}{E_\Sigma - E_\Pi} = \frac{\eta}{2}(\boldsymbol{\sigma}_x \mp i\boldsymbol{\sigma}_y) \tag{C.2}$$

The expressions for the $\hat{H}_{OZ}$ and $\hat{H}_{SO}$ operator matrix elements are chosen in agreement with the notation used by Bolvin[20] in order to facilitate a direct comparison between the g tensor shift formulas obtained in ref 20 for linear molecules. For the nonrelativistic g tensor, $\Delta\vec{\vec{g}}_{SO}$ gives a nonvanishing contribution only to the perpendicular g tensor shift component $\Delta g_\perp$ (see eqs 13 and 14)

$$\Delta g_\perp = [\Lambda_{SO}]_{xx} = 4\left(\frac{[^{\Sigma\Pi_+}\Omega_{OZ}]_x[^{\Pi_+\Sigma}\Omega_{SO}]_x}{E_\Sigma - E_\Pi} + \frac{[^{\Sigma\Pi_-}\Omega_{OZ}]_x[^{\Pi_-\Sigma}\Omega_{SO}]_x}{E_\Sigma - E_\Pi}\right) = 4\eta L \tag{C.3}$$

Therefore, in agreement with the conventional g tensor formalism, our two-state model predicts that only the perpendicular component of the nonrelativistic g tensor of linear molecules deviates significantly from the free-electron g factor. The leading-order relativistic corrections to the electronic g tensor shift are obtained from the remaining contributions to $\Delta\vec{\vec{g}}$ considered in this two-state model, namely $\Delta\vec{\vec{g}}_{SO/SR}$, $\Delta\vec{\vec{g}}_{SO/SO}$, and $\Delta\vec{\vec{g}}_{SZ/SO}$.

Let us first consider the spin−orbit induced contributions to the g tensor shift of $\mathcal{O}(\alpha^4)$. One of these contributions is the pure spin−orbit correction to the electronic Zeeman effect, i.e., $\Delta\vec{\vec{g}}_{SZ/SO}$, which gives contributions to both the parallel and perpendicular components of the electronic g tensor shift, according to eqs 19 and 24

$$\Delta g_\perp = 2[\Sigma_{SZ/SO}]_{xx} = -4\left(\frac{[^{\Sigma\Pi_+}\Omega_{SO}]_y[^{\Pi_+\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2} + \frac{[^{\Sigma\Pi_-}\Omega_{SO}]_y[^{\Pi_-\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2}\right) = -2\eta^2 \tag{C.4}$$

$$\Delta g_{||} = 2[\Sigma_{SZ/SO}]_{zz} = -4\left(\frac{[^{\Sigma\Pi_+}\Omega_{SO}]_x[^{\Pi_+\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} + \frac{[^{\Sigma\Pi_-}\Omega_{SO}]_x[^{\Pi_-\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} + \frac{[^{\Sigma\Pi_+}\Omega_{SO}]_y[^{\Pi_+\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2} + \frac{[^{\Sigma\Pi_-}\Omega_{SO}]_y[^{\Pi_-\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2}\right) = -4\eta^2 \tag{C.5}$$

whereas the other contribution is the higher-order spin−orbit correction to the orbital Zeeman effect, i.e., $\Delta\vec{\vec{g}}_{SO/SO}$ and affects only the parallel component of the g tensor (see eqs 23 and 24)

$$\Delta g_{\parallel} = [\Lambda_{SO/SO}]_{zz} = 2i \sum_{j,j'=x,y,z} \epsilon_{jj'z} \frac{[{}^{\Sigma\Pi_{(+,-)}}\Omega_{SO}]_j [{}^{\Pi_{(+,-)}\Pi_{(+,-)}}\Omega_{OZ}]_z [{}^{\Pi_{(+,-)}\Sigma}\Omega_{SO}]_{j'}}{(E_\Sigma - E_\Pi)^2} = 2i\left( \frac{[{}^{\Sigma\Pi_+}\Omega_{SO}]_x [{}^{\Pi_+\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2} - \frac{[{}^{\Sigma\Pi_+}\Omega_{SO}]_y [{}^{\Pi_+\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} - \frac{[{}^{\Sigma\Pi_-}\Omega_{SO}]_x [{}^{\Pi_-\Sigma}\Omega_{SO}]_y}{(E_\Sigma - E_\Pi)^2} + \frac{[{}^{\Sigma\Pi_-}\Omega_{SO}]_y [{}^{\Pi_-\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} \right) = 2\eta^2 \quad \text{(C.6)}$$

We have here explicitly inserted the values of the reduced orbital Zeeman effect operators between the complex components of the $|^2\Pi\rangle$ state, which are equal to $[{}^{\Pi_\pm\Pi_\pm}\Omega_{OZ}]_z = \pm 1$ and $[{}^{\Pi_\pm\Pi_\mp}\Omega_{OZ}]_z = 0$, respectively. This selection of values for the reduced orbital Zeeman effect operator matrix elements between the complex $^2\Pi$ state components is well justified for a $|^2\Pi\rangle$ state dominated by pure $p_x$ and $p_y$ atomic orbitals (see ref 20), and according to our quadratic response double residue calculations of $[{}^{\Pi_{(+,-)}\Pi_{(+,-)}}\Omega_{OZ}]_z$ type matrix elements holds very well for all investigated dihalogen anion radicals. Collecting these contributions to the electronic g tensor shift, which represent the spin−orbit effects on the g tensor, we obtain the following expression for the perpendicular and parallel components of $\Delta\vec{\vec{g}}$ in linear molecules

$$\Delta g_\perp = 4\eta L - 2\eta^2 \quad \text{and} \quad \Delta g_{\parallel} = -2\eta^2 \quad \text{(C.7)}$$

which are equivalent to those obtained by Bolvin (see eq 39) in ref 20). Therefore, the two-state model presented here, which deals with spin−orbit interaction-induced contributions, is identical to the "two-step" approach developed by Bolvin[20] if a single $|^2\Pi\rangle$ state is used in both models for linear molecules, and contributions from quartet states are neglected in the "two-step" approach.

Finally, let us examine another aspect of the two-state model, namely the treatment of the passive scalar-relativistic effects, which are accounted for by the inclusion of the $\Delta\vec{\vec{g}}_{SO/SR}$ term in the g tensor shift evaluation formulas. The $\Delta\vec{\vec{g}}_{SO/SR}$ only affects the perpendicular component of the g tensor shift, as this contribution corrects $\Delta\vec{\vec{g}}_{SO}$, and gives rise to the following contribution according to eqs 22 and 24):

$$\Delta g_\perp = [\Lambda_{SO/SR}]_{xx} = 4\left( \frac{{}^{\Pi_+\Pi_+}\Omega_{SR} [{}^{\Sigma\Pi_+}\Omega_{OZ}]_x [{}^{\Pi_+\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} + \frac{{}^{\Pi_-\Pi_-}\Omega_{SR} [{}^{\Sigma\Pi_-}\Omega_{OZ}]_x [{}^{\Pi_-\Sigma}[\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} - \frac{{}^{\Sigma\Sigma}\Omega_{SR} [{}^{\Sigma\Pi_+}\Omega_{OZ}]_x [{}^{\Pi_+\Sigma}\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} + \frac{{}^{\Sigma\Sigma}\Omega_{SR} [{}^{\Sigma\Pi_-}[\Omega_{OZ}]_x [{}^{\Pi_-\Sigma}[\Omega_{SO}]_x}{(E_\Sigma - E_\Pi)^2} \right) = 4\eta L\zeta \quad \text{(C.8)}$$

where we have introduced the energy-weighted reduced matrix element of the $\hat{H}_{SR}$ operator

$$\zeta = \frac{{}^{\Pi_\pm\Pi_\pm}\Omega_{SR} - {}^{\Sigma\Sigma}\Omega_{SR}}{E_\Sigma - E_\Pi} \quad \text{(C.9)}$$

The passive scalar-relativistic correction is thus expected to be of similar magnitude as the higher-order spin−orbit correction to $\Delta g_\perp$, i.e., $4L\eta\zeta \approx 2\eta^2$ and this approximate equality can be expected to hold for most linear molecules. Combining the scalar-relativistic contribution to $\Delta\vec{\vec{g}}$ with the previously described spin−orbit contribution, we obtain the final expression for evaluating electronic g tensor shifts of linear molecules in the two-state model:

$$\Delta g_\perp = 4\eta L(1 + \zeta) - 2\eta^2 \quad \text{and} \quad \Delta g_{\parallel} = -2\eta^2 \quad \text{(C.10)}$$

It is clear that the scalar-relativistic and spin−orbit corrections to $\Delta g_\perp$ almost cancel each other and that $\Delta g_\perp$ becomes roughly equal to the nonrelativistic g tensor shift value, i.e., $\Delta g_\perp \approx 4\eta L$. In view of these findings, we emphasize that scalar relativistic effects only play a minor role in the electronic g tensor in linear molecules and that the spin−orbit interaction is responsible for the large negative parallel g tensor shift in linear molecules containing heavy elements. A similar conclusion about the importance of the spin−orbit contribution to $\Delta g_{\parallel}$ has previously been obtained both theoretically and numerically by Bolvin,[20] as well as observed by Malkin et al. in their density-functional, two-component Douglas−Kroll calculations on the dihalogen anion radicals.[18]

Finally, before concluding the discussion of the two-state model for evaluating electronic g tensors in linear molecules with $|^2\Sigma\rangle$ type ground state, we would like to validate the approach for the dihalogen anion radicals. The electronic g tensor shift computed using the two-state model and the full DGPT treatment according to eq 25 are tabulated in Table 5. The results show that the two-state model results deviate from the ones obtained using a full DGPT treatment by on average 10%, and the largest deviation is observed for BLYP in the $I_2^-$ radical anion. On the basis of this limited test example set, we can recommend the two-state model for estimating electronic g tensors in linear molecules, since these g tensors are dominated by the contribution from the singly excited $|^2\Pi\rangle$ state. Furthermore, this model can easily be generalized to a multistate model following the derivation above if the linear molecule features several $|^2\Pi\rangle$ states with significant contributions to the g tensor shift. We believe the two-state model will be useful in the analysis of EPR spectra of linear molecules and help quantify different effects responsible for the g tensor shift in molecules of this kind.

## References


(1) Harriman, J. E. In *Theoretical Foundations of Electron Spin Resonance*; Academic Press: New York, 1978; Chapter 2, pp 111−162.

(2) Kutzelnigg, W. In *Calculations of NMR and EPR Parameters*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH Verlag: Weinheim, Germany, 2004; Chapter 5, pp 43−84.

(3) Neese, F. *Curr. Opin. Chem. Biol.* **2003**, *7*, 125–135.

(4) Lushington, G. H.; Grein, F. *Theor. Chem. Acc.* **1996**, *93*, 259–267.

(5) Lushington, G. H.; Grein, F. *J. Chem. Phys.* **1997**, *106*, 3292.

(6) Vahtras, O; Minaev, B.; Ågren, H. *Chem. Phys. Lett.* **1997**, *281*, 186–192.

(7) Schreckenbach, G.; Ziegler, T. *J. Phys. Chem. A* **1997**, *101*, 3388–3399.

(8) van Lenthe, E; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **1997**, *107*, 2488.

(9) Jayatilaka, D. *J. Chem. Phys.* **1998**, *108*, 7587.

(10) Malkina, O. L.; Vaara, J.; Schimmelpfennig, B.; Munzarova, M.; Malkin, V. G.; Kaupp, M. *J. Am. Chem. Soc.* **2000**, *122*, 9206–9218.

(11) Lushington, G. H. *J. Phys. Chem. A* **2000**, *104*, 2969–2974.

(12) Neese, F. *J. Chem. Phys.* **2001**, *115*, 11080.

(13) Neyman, K. M.; Ganyushin, D. I.; Matveev, A. V.; Nasluzov, V. A. *J. Phys. Chem. A* **2002**, *106*, 5022–5030.

(14) Kaupp, M.; Reviakine, R.; Malkina, O. L.; Arbuznikov, A.; Schimmelpfennig, B.; Malkin, V. G. *J. Comput. Chem.* **2002**, *23*, 794–803.

(15) Brownridge, S; Grein, F; Tatchen, J.; Kleinschmidt, M.; Marian, C. M. *J. Chem. Phys.* **2003**, *118*, 9552.

(16) Rinkevicius, Z.; Telyatnyk, L.; Sałek, P.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2003**, *119*, 10489.

(17) Manninen, P.; Vaara, J.; Ruud, K. *J. Chem. Phys.* **2004**, *121*, 1258.

(18) Malkin, I.; Malkina, O. L.; Malkin, V. G.; Kaupp, M. *J. Chem. Phys.* **2005**, *123*, 244103.

(19) Komorovský, S.; Repiský, M.; Malkina, O. L.; Malkin, V. G.; Malkin, I.; Kaupp, M. *J. Chem. Phys.* **2006**, *124*, 84108.

(20) Bolvin, H. *Chem. Phys. Chem.* **2006**, *7*, 1575–1589.

(21) Arratia-Pérez, R.; Hernandez-Acevedo, L.; Malli, G. L. *J. Chem. Phys.* **2004**, *121*, 7743.

(22) Arratia-Pérez, R.; Malli, G. L. *J. Chem. Phys.* **2006**, *124*, 74321.

(23) Vancoille, S.; Malmqvist, P. Å.; Pierloot, K. *Chem. Phys. Chem.* **2007**, *8*, 1803–1815.

(24) Fritscher, J.; Hrobárik, P.; Kaupp, M. *J. Phys. Chem. B* **2007**, *111*, 4616–4629.

(25) Fritscher, J.; Hrobárik, P.; Kaupp, M. *Inorg. Chem.* **2007**, *46*, 8146–8161.

(26) Bolvin, H. *Inorg. Chem.* **2007**, *46*, 417–427.

(27) Chang, Ch.; Pelissier, M.; Durand, Ph. *Phys. Scr.* **1986**, *34*, 394.

(28) Heully, J. L.; Lindgren, I.; Lindroth, E.; Lundquist, S.; Mårtensson-Pendrill, A. M. *J. Phys. B* **1986**, *19*, 2799.

(29) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.

(30) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 1272.

(31) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89.

(32) Stein, M.; van Lenthe, E.; Baerends, E. J.; Lubitz, W. *J. Phys. Chem. A* **2001**, *105*, 416–425.

(33) Saladino, C. A.; Larsen, C. S. *J. Phys. Chem. A* **2003**, *107*, 1872–1878.

(34) Pietrzyk, P; Sojka, Z. *J. Phys. Chem. A* **2005**, *109*, 10571–10581.

(35) Rinkevicius, Z.; Tunell, I.; Sałek, P.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2003**, *119*, 34.

(36) Rinkevicius, Z.; Jha, P. C.; Oprea, C. I.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2007**, *127*, 114101.

(37) Abragam, A.; Bleaney, B. In *Electronic Paramagnetic Resonance of Transition Ions*; Clarendon Press: Oxford, UK, 1970; Chapter 5, pp 652−653.

(38) Patchkovskii, S.; Strong, R. T.; Pickard, C. J.; Un, S. *J. Chem. Phys.* **2005**, *122*, 214101.

(39) Hess, B. A.; Marian, C. M.; Wahlgren, U.; Gropen, O. *Chem. Phys. Lett.* **1996**, *251*, 365–371.

(40) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1210. Parameterization V.

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

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

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

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

(45) DALTON, a molecular electronic structure program, Release 2.0 (2005): http://www.kjemi.uio.no/software/dalton/dalton.html.

(46) Bruna, P. J.; Grein, F. *Chem. Phys.* **1999**, *249*, 169–182.

(47) Bruna, P. J.; Grein, F. *Int. J. Quantum Chem.* **2000**, *77*, 324–335.

(48) Fernandez, B.; Jørgensen, P.; Byberg, J.; Olsen, J.; Helgaker, T.; Jensen, H. J. A. *J. Chem. Phys.* **1992**, *97*, 3412.

(49) Patchkovskii, S.; Schreckenbach, G. In *Calculation of NMR and EPR Parameters: Theory and Applications*, 1st ed; Kaupp, M., Bühl, M., Malkin, V. G., Eds; Wiley-VCH: Weinheim, Germany, 2004; pp 505−532.

(50) Knight, L. B.; Earl, E.; Ligon, A. R.; Cobranchi, D. P. *J. Chem. Phys.* **1986**, *85*, 1228.

(51) Weltner, Jr. W. In *Magnetic Atoms and Molecules*; Dover: New York, 1990; pp 34−35.

(52) Löwdin, P. O.; Goscinski, O. *Int. J. Quantum Chem.* **1971**, *5*, 685–705.

(53) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.

(54) Neese, F.; Solomon, E. I. *Inorg. Chem.* **1998**, *37*, 6568–6582.

CT800053F

JCTC Journal of Chemical Theory and Computation

# Benzene Dimer: High-Level Wave Function and Density Functional Theory Calculations

M. Pitoňák,[†] P. Neogrády,[‡] J. Řezáč,[†] P. Jurečka,[†,§] M. Urban,[‡] and P. Hobza*[,†,§]

*Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i., Flemingovo nám. 2, 166 10 Praha 6, Czech Republic, Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava 4, Slovak Republic, and Department of Physical Chemistry, Palacky University, tr. Svobody 26, 771 46, Olomouc, Czech Republic*

Received June 19, 2008

**Abstract:** High-level OVOS (optimized virtual orbital space) CCSD(T) interaction energy calculations (up to the aug-cc-pVQZ basis set) and various extrapolations toward the complete basis set (CBS) limit are presented for the most important structures on the benzene dimer potential energy surface. The geometries of these structures were obtained via an all-coordinate gradient geometry optimization using the DFT-D/BLYP method, covering the empirical dispersion correction fitted exclusively for this system. The fit was carried out against two estimated CCSD(T)/CBS potential energy curves corresponding to the distance variation between two benzene rings for the parallel-displaced (PD) and T-shaped (T) structures. The effect of the connected quadruple excitations on the interaction energy was estimated using the CCSD($TQ_f$) method in a 6-31G*(0.25) basis set, destabilizing the T and T-shaped tilted (TT) structures by ≈0.02 kcal/mol and the PD structure by ≈0.04 kcal/mol. Our best CCSD(T)/CBS results show, within the error bars of the applied methodology, that the energetically lowest-lying structure is the TT structure, which is nearly 0.1 kcal/mol more stable than the almost isoenergetic PD and T structures. The specifically parametrized DFT-D/BLYP method leads to a correct energy ordering of the structures, with the errors being smaller by 0.2 kcal/mol with respect to the most accurate CCSD(T) values.

## Theoretical Background

Throughout the history of computational chemistry, the benzene dimer has been quite extensively studied system for several reasons. First, it is an ideal example of a system bound by $\pi-\pi$ interaction, which is in numerous cases present in many biologically relevant systems (stacking interactions in DNA base pairs, aromatic side-chain interactions in proteins) as well as in material and nanoscience.[1] Second, an accurate description of the $\pi-\pi$ interaction is in general a challenging task for the methods of quantum chemistry.[2] Noncovalent interaction between benzene rings results in a very shallow potential energy surface (PES), where it seems even more problematic to obtain the proper geometry of the global and local minima, saddle points, and barriers than to obtain the accurate, single-point stabilization energies themselves.[3] Even the experimentally known values of the stabilization energies of the benzene dimer[4,5] have rather large error bars, and their correspondence to particular structures has not yet been completely resolved.

The goal of this paper is to provide a benchmark calculation on both the geometry and the electronic component of the stabilization energy of the most important structures on the benzene dimer PES, which will make it possible to identify the "global" minimum and a proper energy ordering of the other important structures.

* Corresponding author. E-mail: pavel.hobza@uochb.cas.cz.
[†] Academy of Sciences of the Czech Republic.
[‡] Comenius University.
[§] Palacky University.

10.1021/ct800229h CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

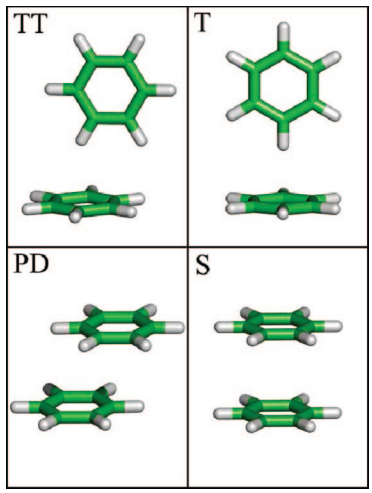



***Figure 1.*** Investigated structures of the benzene dimer: T-shaped tilted $c_s$ (TT), T-shaped $c_{2v}$ (T), parallel-displaced $c_{2h}$ (PD), and sandwich $d_{6h}$ (S).

We are certainly aware of the fact that the transition barriers separating the global and local minima are quite low[3] (of an order of magnitude of 0.1 kcal/mol) and that they are likely to disappear once the zero-point vibrational energy is included. This means that the structures populated at various temperatures can be quite different from the minima localized at the PES (see the proceeding paper[6]). An accurate description of the PES is, however, important in order to be able to understand the nature of stabilization in various structures. It must be kept in mind that these structures (T-shaped tilted $c_s$ (TT), T-shaped $c_{2v}$ (T), parallel-displaced $c_{2h}$ (PD), and sandwich $d_{6h}$ (S); see Figure 1) represent the model motifs occurring in biomacromolecules and complex molecular systems. In these cases, the energy barriers will be much higher and might lead to the coexistence of various structures. A dominant role in the stabilization of all benzene dimer structures is played by dispersion energy. Besides this energy, the T-shaped and PD structures are stabilized also by the electrostatic quadrupole—quadrupole term, and the former structure is stabilized additionally by the H-bonding contribution (in this case the blue-shifting H-bond[7]). In the case of the S structure, the electrostatic energy is repulsive. Both the electrostatic and H-bonding terms are properly described already by such low-level quantum mechanical (QM) methods as Hartree—Fock (HF) or density functional theory (DFT) methods. On the other hand, the description of dispersion energy requires the use of the most advanced QM procedures. Within the wave function (WF) theory, the description of the dispersion energy is notorious for its slow convergence with respect to both the level of the applied theory and atomic orbital (AO) basis set size. In the case of the benzene[8] and uracil[9] dimer an "oscillatory" convergence of the stabilization energy, manifested by a strong overestimation at the Møller—Plesset second-order perturbation theory (MP2) level and underestimation at the MP3 or CCSD (coupled clusters with single and double excitations) level, is a good example. CCSD energies corrected for the effect of the connected triple excitation treated in either the fully iterative or perturbative way are in good agreement[10] and bring the results into reasonable agreement with the experiment. However, the question of the importance of the higher connected excitation in the benzene dimer interaction was raised in the past[11] but due to the computationally prohibitive size of the system could not be convincingly answered. Calculations on smaller, model systems suggest that the effect could be of an order of magnitude of tenths of kilocalories per mole. Quadruple (and higher) excitations typically gain in importance in cases where the single reference WF begins to acquire a multireference character. This is most likely not the case of dominantly dispersion-bound complexes, where the interaction on the orbital level is practically negligible. The true source of the importance of quadruples originates in the incompleteness of the (in our case) CCSD(T) supermolecular interaction energy, as pointed out by Jeziorski et al.[12]

Another important problem concerns the structures of the dimer. The most investigated structures have been the T, S, and PD structures, with their geometries being determined by point-by-point energy optimization.[13] However, it has recently[1,14] been shown that the T structure does not correspond to an energy minimum but is a transition structure. The minimum has a tilted T structure (TT) with lower symmetry ($c_s$ instead of $c_{2v}$). In this case, the point-by-point optimization becomes difficult and more accurate geometries are provided by a gradient optimization.

The most accurate calculations on the benzene dimer published so far were carried out by Janowski and Pulay.[13] Their results calculated at the QCISD(T)|CBS level (Helgaker's extrapolation[15] from Dunning-type aug-cc-pVTZ and aug-cc-pVQZ basis sets)[16] using geometries from the QCISD(T)|aug-cc-pVTZ optimization of the intermolecular distances have led to the following ordering of the stabilization energies: T(2.68) ≥ PD(2.66) > S(1.65) (the stabilization energies in parentheses, in kilocalories per mole). Further publications, suggesting that the most stable benzene dimer structure is TT, have recently been issued by Distasio et al.[14] and Lee et al.,[1] in the first of which the stabilization energy was calculated at the estimated CCSD(T)|complete basis set (CBS) level (MP2|CBS + dCCSD(T)|6-311+G(2df,p), see below) leading to the ordering: TT(2.66) > T(2.54) > PD(2.27). The authors of the latter publication, apart from presenting an extensive literature review on this topic, suggested the ordering: TT(2.84/2.84) > T(-/2.77) > PD(2.73/2.62) > S(1.66/1.53). The first value in the parentheses was obtained as both a CCSD(T)|CBS (MP2|CBS + dCCSD(T)|aug-cc-pVTZ) geometry optimization and single-point calculation, whereas the second value was obtained as a CCSD(T)|CBS (with the same methodology) on the Boys—Bernardi[17] counterpoise basis set superposition error (BSSE)-corrected RI-MP2|aug-cc-pVDZ optimized geometry. The latter results are almost identical with those published recently by Bludský et al.,[18] who introduced DFT/CCSD(T) correction scheme and applied it for the geometry optimization. Upon the basis of these geometries, they estimated the CCSD(T)|CBS as a Helgaker's extrapolation of the spin-component-scaled MP2

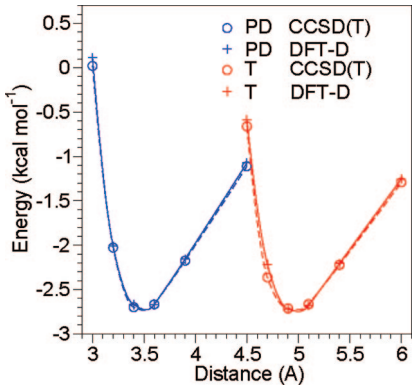



***Figure 2.*** Fitting of the dispersion term for the parallel displaced (PD, blue/left) and T-shaped (T, red/right) benzene dimer; (circles) reference CCSD(T)|CBS values, (lines and crosses) fitted DFT-D energy.

(SCS-MP2)[19] energies from the aug-cc-pVTZ and aug-cc-pVQZ basis sets, corrected with the dCCSD(T)|aug-cc-pVDZ.

In the paper being presented, we will investigate such benzene dimer structures whose geometries have been determined by the all-coordinate gradient optimization based on the DFT-D procedure, which closely mimics the estimated CCSD(T)|CBS results. The energies of the dimer were determined at the highest affordable CCSD(T)|CBS level with various extrapolations adopted.

## Methodology

The accurate geometry of the conformers investigated plays an important role in obtaining highly accurate stabilization energies. Since the CCSD(T) gradient optimization for benzene dimer is impractical, typically the intrasystem coordinates are kept frozen, and step-by-step one-dimensional optimization is applied.

Our approach is based on two key elements: (a) the DFT-D/BLYP|TZVP[20] procedure containing an empirical dispersion term, which was parametrized specifically for the benzene dimer; the resulting DFT-D method was applied for an all-coordinate geometry optimization; (b) highly accurate single-point energies obtained by BSSE corrected CCSD(T), calculated in truncated optimized virtual orbital space (OVOS).[21]

The OVOS method is a tool for accelerating calculations of the correlation energy (MP2, MP3,..., CCSD, CCSD(T)) via expanding the correlated WF in a subset of unitary transformed HF virtual orbitals (VOs). Optimal unitary transformation is found by maximizing the overlap of the first-order many-body perturbation (MBPT) WF in the full and the truncated virtual orbital space (VOS). For instance, CCSD(T) in OVOS truncated to 70% is the CCSD(T) energy calculated in the unitary transformed HF VOs, from which only lowest 70% were correlated. This energy will also be referred to as the CCSD(T)$^{\mathrm{OVOS}}$ 70%.

The dispersion correction in the DFT-D procedure was reoptimized to reproduce the highly accurate CCSD(T)|CBS reference data. Six reference points were calculated for both the T and PD structures; see Figure 2. The BSSE corrected CCSD(T)|CBS interaction energies were constructed as a sum of the HF|aQZ interaction energies (aXZ refers to Dunning's aug-cc-pVXZ basis set),[16] the MP2|CBS term (Helgaker's extrapolation;[15] see eq 1 from aTZ and aQZ basis sets) and the dCCSD(T)$^{\mathrm{OVOS}}$ correction term (i.e., the difference between MP2 and CCSD(T) interaction energies determined with an aTZ basis set in OVOS truncated to 70%). All three parameters in the dispersion formula, i.e. the scaling factor for the van der Waals radii, $s_R$, the global scaling of the dispersion energy according to Grimme,[22] $s_6$, and the exponent of the damping function, $\alpha$, were included in the optimization. For the sake of computational speed, we opted for the B-LYP functional and the relatively small, TZVP basis set. As the main guideline for the fitting, we used the unsigned error of the interaction energy, weighted by the Boltzmann factor (at 50 K), in order to ensure that the energy minimum was described with a higher accuracy than the rest of the PES. Some arbitrariness entered the fitting procedure through the compromise between the minimum weighted error and the requirement that both conformers have the correct relative energy in their minima. The resulting curves corresponding to the parameter set $s_R = 0.88$, $s_6 = 1.503$, and $\alpha = 6$ are compared to the reference data in Figure 2.The agreement between the reference CCSD(T)|CBS and DFT-D values is very good, but note that this is on account of the highly customized dispersion parameters, which differ significantly from the transferable parameters given in Jurečka et al.[20]

The DFT-D procedure, connected with the all-coordinate gradient optimization, yielded optimal TT, T, PD, and S structures (see section 3.1). For these structures, single-point CCSD(T) or OVOS CCSD(T)$_2$ (CCSD(T)$_2$ = MP2$^{\mathrm{FullVOS}}$ + dCCSD(T)$^{\mathrm{OVOS}}$) energies were calculated in the aDZ, the aTZ with the OVOS truncated to 70%, the aQZ with the OVOS truncated to 60%, and the MP2 energies calculated in a5Z. These rather demanding calculations were carried out using a new, parallel CCSD(T) code[23] based on Cholesky decomposed two-electron integrals, implemented in the MOLCAS 7[24] program package. In all of the calculations, the $10^{-5}$ threshold was used for the Cholesky decomposition of the two-electron integrals. In the aQZ and a5Z basis set calculations, the $10^{-5}$ threshold for eliminating the linear-dependent basis functions in the HF-SCF step resulted in the deletion of 50−56 basis functions in the aQZ and 103−114 basis functions in the a5Z basis sets, depending on the particular structure. Those calculations of the stabilization energies using these sets of thresholds lead to errors not larger than 0.03 kcal/mol, which is within the error bars of the CCSD(T) calculations in the truncated OVOS.[25]

In this work, two CBS extrapolation schemes have been used: (a) Helgaker's scheme[15]

$$E_X^{\mathrm{corr}} = E_{\mathrm{CBS}}^{\mathrm{corr}} + k/X^3 \qquad (1)$$

where $E_{\mathrm{CBS}}^{\mathrm{corr}}$ is an estimate of the CBS correlation energy and $k$ is an arbitrary constant; (b) Kim's scheme,[1,27] using both BSSE corrected ($\Delta E_X^{\mathrm{b}}$) and uncorrected ($\Delta E_X^{n}$) interaction energies

$$\delta_X = \Delta E_X^{\mathrm{b}} - \Delta E_X^{n} \qquad (2)$$

$$\varepsilon_X = \Delta E_X^{\mathrm{b}} + \Delta E_X^{n} \quad (3)$$

$$\Delta E_{\mathrm{CBS}} = 1/2(\delta_X \varepsilon_{X+1} - \delta_{X+1}\varepsilon_X)/(\delta_X - \delta_{X+1}) \quad (4)$$

in the (typically a Dunning-type) basis set with cardinality $X$. Use of this pseudointerpolation scheme is justified by the fact that the BSSE corrected and BSSE uncorrected interaction energies converge with increasing of the basis set size to the same CBS value and an observation that (especially the correlation part) the BSSE corrected and BSSE uncorrected values approach the CBS interaction energy from each side.[26] Furthermore, this "extrapolation" does not depend on any external parameter(s), contrary to the Helgaker's scheme and its variants, and is more indulgent to the numerical noise caused by superposition of the several approximations made in our computational scheme.

## 3. Results and Conclusions

**3.1. Geometries Obtained from the DFT-D Optimization.** The coordinates of all the optimized structures are summarized in the Supporting Information, and their main internal geometry parameters are presented in Table 1.

**3.2. Role of the connected quadruple excitations.** The most relevant work on the importance of the connected quadruple excitations in a coupled-cluster framework[11] was done on the acetylene dimer with the CCSD(TQ$_f$)[28,29] method using the aug-cc-pVDZ basis set, where the largest contribution from quadruples was 0.04 kcal/mol for the T-shaped structure. For larger systems, calculations were feasible only in the 6-31G*(0.25)[30] basis set and the authors obtained quite large contributions, i.e. 0.13 kcal/mol for the butadiene in an antiparallel stacked geometry and 0.20 kcal/mol for the furan dimer in the $C_{2h}$ stacked geometry. If similar contributions from the quadruples would arise in the benzene dimer, it could completely change their energy ordering. CCSD(TQ$_f$) calculation on our optimized benzene dimer structures were carried out using the ACES II program package,[31] and the results are shown in Table 2.The accuracy of these results obtained with such a small basis set is, however, questionable. Two effects, directed against each other, might play a role. It is known that the higher-order connected excitations in the coupled-cluster wave operator converge faster with basis set size, but on the other hand, their contribution in small basis sets tends to be artificially large as they try to "imitate" the basis set effect. By looking at the rate of convergence of the CCSD-MP2 and the (T) contributions to the stabilization energy for 6-31G*(0.25) and aug-cc-pVXZ basis sets for the benzene dimer structures under consideration, we believe that the numbers in Table 2 might still be a reasonable estimate of their CBS values, especially in the sense of their relative importance for the different structures. Repulsion is in case of the stacked structures (PD and S) somewhat stronger than in case of the T-shaped structures (TT, T), thus supporting our conclusions (see the next section) that the TT structure is likely to be the most stable one.

***Table 1.*** Geometry Parameters [Å] of the Investigated Benzene Dimer Structures

| structure | lateral disp[a] | distance |
|---|---|---|
| T | | 4.948 |
| TT | 0.875[b] | 4.828 |
| S | | 3.847 |
| PD | 1.768[c] | 3.486 |

[a] Displacement with respect to the center of mass of the individual benzene rings. [b] Benzene ring in the plane of the $C_2$ axis of the "T" structure is rotated clockwise against the axis by 24.07° (see the TT and T structures in Figure 1). [c] Horizontal displacement is in the direction of the C−H bond (unlike the horizontal displacement perpendicular to the C−C bond in Janowski et al.[13] and Sinnokrot et al.[32,33]

***Table 2.*** Contribution from the Connected Quadruple Excitation (kcal/mol) to the Stabilization Energy Obtained by the CCSD(TQ$_f$)|6-31G*(0.25) Method|Basis Set for the Benzene Dimer Structures Shown in Figure 1[a]

| structure | $\Delta E$(Q$_f$) |
|---|---|
| PD | 0.043 |
| TT | 0.024 |
| T | 0.021 |
| S | 0.038 |

[a] All contributions are repulsive.

***Table 3.*** Benzene Dimer Stabilization Energies [kcal/mol] Obtained for Various Structures, Basis Sets, Methods, and CBS Extrapolations[i]

| method\|basis set | PD | PD$_p$ | TT | T | S |
|---|---|---|---|---|---|
| CCSD(T)\|D | 2.15 | 2.29 (2.30)[a] | 2.44 | 2.28[b] | 1.27[c] |
| CCSD(T)$_2$\|T 70% | 2.49 | 2.53 (2.55) | 2.66 | 2.57 | 1.51 |
| CCSD(T)$_2$\|Q 60% | 2.63 | 2.64 (2.61) | 2.75 | 2.65 | 1.61 |
| (T → Q)[d] | 2.73 | 2.72 (2.66) | 2.81 | 2.71 | 1.68 |
| (T → Q)[e] | 2.66 | 2.65 | 2.76 | 2.68 | 1.63 |
| (Q → 5)[f] | 2.74 | 2.73 | 2.81 | 2.71 | 1.68 |
| (Q → 5)[g] | 2.70 | 2.69 | 2.78 | 2.69 | 1.64 |
| DFT-D/BLYP\|TZVP | 2.88 | 2.81 | 2.93 | 2.80 | 1.84 |
| DFT-D[h] /BLYP\|TZVP | 2.57 | 2.51 | 2.33 | 2.03 | 1.45 |

[a] QCISD(T) stabilization energies by Janowski et al.[13] [b] Corresponding QCISD(T)|$X$ and CBS stabilization energies,[13] obtained in a slightly different, QCISD(T)|T-optimized geometry, are 2.30, 2.57, 2.64, and 2.68 kcal/mol. [c] Corresponding QCISD(T)|$X$ and CBS stabilization energies,[13] obtained in a slightly different, QCISD(T)|T-optimized geometry, are 1.40, 1.60, 1.64, and 1.65 kcal/mol. [d] (T → Q) extrapolation using Helgaker's formula in the form: SCF|Q + corr CCSD(T)$_2$| (T → Q). [e] Kim's extrapolation scheme using both BSSE corrected and uncorrected CCSD(T)$_2$ interaction energies in T, Q. [f] (Q → 5) extrapolation using Helgaker's formula in the form: SCF|5 + corr MP2|(Q → 5) + dCCSD(T)$^{\mathrm{OVOS}}$|(T → Q). [g] Kim's scheme with interaction energies calculated as SCF|$X$ + corr MP2|$X$ + dCCSD(T)$^{\mathrm{OVOS}}$|($X$ − 1), where $X$ stands for Q, 5 and $X$ − 1 for T, Q. [h] Empirical dispersion term with the original parameters fitted to S22 test set.[20] [i] $X$ = T, Q, 5 stand for Dunning's aug-cc-pVXZ basis sets.

**3.3. Stabilization Energies of the Dimers.** The stabilization energies obtained for the optimized structures in Figure 1 are shown in Table 3.

This table provides DFT-D and CCSD(T) energies as well as CBS extrapolations for the structures obtained as described above and the "PD$_p$", the one selected structure optimized by Janowski et al.,[13] which is equivalent to our PD structure. The PD$_p$ structure serves as a reference for our methodology of geometry optimization, for the accuracy of the single-point calculations and extrapolations. Let us first to compare our results obtained for the PD structure with the values obtained for the PD$_p$ structure, which was point-by-point optimized (with the geometry of the monomers being fixed)

by means of the expensive BSSE corrected QCISD(T)|aTZ method.[13] So as to provide an estimate of errors of our single-point CCSD(T) energies in the truncated OVOS, we present our values along with their QCISD(T) stabilization energies (in parentheses). The difference between the CCSD(T) and QCISD(T) stabilization energies in this particular case was shown to be very small, as roughly 0.004 kcal/mol overestimated by QCISD(T). The errors with respect to our OVOS CCSD(T) values in the $PD_p$ geometry are −0.01 in the aDZ basis set, −0.02 in the aTZ basis set and 0.03 kcal/mol in the aQZ basis set. An important fact to be emphasized is that our aTZ calculation was carried out with 534 VOs instead of 786 VOs and our aQZ calculation with only 851 VOs instead of 1416 VOs (after the deletion of 54 basis functions due to their linear dependence), which resulted in an almost 1 order of magnitude acceleration in the CPU time. The stabilization energies in aDZ and aTZ slightly underestimate the reference values due to the error introduced by the truncation of the OVOS. The error in the aQZ basis set has the opposite sign as a consequence of the superposition of the errors of the truncation of the OVOS and the error of the elimination of the linear dependent basis functions (which itself has an effect of ≈0.01 kcal/mol). Such an unbalanced error distribution for different basis sets can result in error amplification (by up to 0.06 kcal/mol) when Helgaker's extrapolation scheme (see the footnote of Table 3) is used. However, Kim's extrapolation scheme, due to its more "interpolative" character, seems to be less sensitive, and agrees with the reference data, being within 0.02 kcal/mol. This is the reason that the extrapolation labeled in the table as $(Q \rightarrow 5)^g$ is, in our opinion, the most reliable and serves as our benchmark data. Concerning the accuracy of the geometries obtained by our parametrized DFT-D/BLYP|TZVP for PD, T, and S structures, our values are within 0.03 kcal/mol error bars on the CCSD(T)$_2$| aQZ 60% OVOS level with respect to the reference values, which is quite surprising for a method as cheap as the DFT-D.

The energy ordering of the structures is preserved while increasing the basis set size. The TT structure is the most stable, separated by ≈0.1 kcal/mol from the other structures in all the basis sets and CBS. On the other hand, the S structure is the least stable, separated from the other structures by almost 1 kcal/mol. The energy difference between the PD and T structures decreases, starting from 0.13 kcal/mol in the aDZ basis set, converging to almost zero (−0.01 kcal/mol) in the CBS. The ordering of the stabilization energies obtained in our calculations is TT > T ≥ PD > S. This is in agreement with the thus-far most accurate values, provided by Janowski et al.,[13] except for the fact that they excluded the "relatively well" separated, most stable TT structure from their considerations. Our fitted DFT-D method leads to the same energy ordering of the structures, with errors being ≈0.2 kcal/mol when compared to our most accurate CCSD(T)|CBS values. DFT-D with the original parameters fitted to S22 test set performs significantly worse and leads to an incorrect ordering of the structures with respect to the energy. Errors for the stacked structures (PD, $PD_p$, S) are of the same magnitude as with our DFT-D optimized structures for the benzene dimer but are substantially larger for the T-shaped structures (0.45 kcal/mol for the TT structures and 0.66 kcal/mol for the T structure). This clearly demonstrates that, when high accuracy is desired, use of any "universal" parameters for the empirical dispersion term is not possible (for further discussion, see the proceeding paper on the benzene dimer[6]).

**Acknowledgment.** This work was supported by grants No. LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports (MSMT) of the Czech Republic and was part of research project No. Z4 055 0506. It was also supported by the Slovak Research and Development Agency (Contract No. APVV-20-018405). The authors wish to acknowledge the support of Praemium Academiae of the Academy of Sciences of the Czech Republic, awarded to P.H. in 2007.

**Supporting Information Available:** Animation of all the optimized structures. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem.* **2007**, *111*, 3446.

(2) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291.

(3) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345.

(4) Krause, H.; Ernstberger, B.; Neusser, H. J. *Chem. Phys. Lett.* **1991**, *184* (5−6), 411.

(5) Grover, J. R.; Walters, E. A.; Hui, E. T. *J. Chem. Phys.* **1987**, *91*, 3233.

(6) Řezác, J.; Hobza, P. **2008**, *4*, 1835–1840.

(7) Hobza, P.; Špirko, V.; Selzle, H. L.; Schlog, E. W. *J. Phys. Chem. A* **1998**, *102*, 2501.

(8) Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *Chem. Phys. Lett.* **2000**, *319*, 547.

(9) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636.

(10) Pittner, J.; Hobza, P. *Chem. Phys. Lett.* **2004**, *390*, 496.

(11) Hopkins, B. W.; Tschumper, G. S. *J. Phys. Chem. A* **2004**, *108*, 2941.

(12) Korona, T.; Moszynski, R.; Jeziorski, B. *Mol. Phys.* **2001**, *100*, 1723.

(13) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27.

(14) Distasio, R. A., Jr.; von Helden, G.; Steele, R. P.; Head-Gordon, M *Chem. Phys. Lett.* **2007**, *437*, 277.

(15) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286* (3−4), 243.

(16) Dunning, T. H.; Peterson, K. A. *J. Chem. Phys.* **2000**, *113* (18), 7799.

(17) Boys, S. F.; Bernardi, F. *Mol. Phys.* **2002**, *100*, 65.

(18) Bludský, O.; Rubeš, M.; Soldán, P.; Nachtigall, P. *J. Chem. Phys.* **2008**, *128* (11), 114102.

(19) Grimme, S. *J. Chem. Phys.* **2003**, *118* (20), 9095.

(20) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28* (2), 555.

(21) Neogrády, P.; Pitoňák, M.; Urban, M. *Mol. Phys.* **2005**, *103* (15−16), 2141.

(22) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

(23) Neogrády, P.; Aquilante, F.; Noga, J.; Pitoňák, M.; Hobza, P.; Urban, M. in preparation.

(24) Karlstrom, G.; Lindh, R.; Malmqvist, P-A.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P-O.; Cossi, M.; Schimmelpfennig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28* (2), 222.

(25) Dedková, P.; Pitoňák, M.; Neogrády, P.; Černušák, I.; Urban, M. *J. Phys. Chem. A* **2008**, *112*, 7115.

(26) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157.

(27) Shin, I.; Park, M.; Min, S. K.; Lee, E. C.; Suh, S. B.; Kim, K. S. *J. Chem. Phys.* **2006**, *125*, 234305.

(28) Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **1998**, *108*, 9221.

(29) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2005**, *122*, 224102.

(30) Hobza, P.; Šponer, J. *Chem. Rev.* **1999**, *99*, 3247.

(31) Stanton, J. F.; Gauss, J.; Watts, J. D.; Nooijen, M.; Oliphant, N.; Perera, S. A.; Szalay, P. G.; Lauderdale, W. J.; Kucharski, S. A.; Gwaltney, S. R.; Beck, S.; Balková, A.; Bernholdt, D. E.; Baeck, K. K.; Rozyczko, P.; Sekino, H.; Hober, C.; Pittner, J.; Bartlett, R. J. *ACES II*; Quantum Theory Project, University of Florida, 2003. Integral packages included are the following: Almöf, J.; Taylor, P. R. *VMOL*; Taylor, P. R. *VPROPS*. Helgaker, T.; Jensen, H. J.; Olsen, Aa.; Jørgensen, J.; Taylor, P. R. *A modifed version of ABACUS integral derivative package*.

(32) Sinnokrot, M. O.; Vallev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887.

(33) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.

CT800229H

JCTC Journal of Chemical Theory and Computation

# Benzene Dimer: Dynamic Structure and Thermodynamics Derived from On-the-Fly ab initio DFT-D Molecular Dynamic Simulations

Jan Řezáč and Pavel Hobza*

*Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, 166 10 Prague 6, Czech Republic*

Received June 19, 2008

**Abstract:** The dynamic nature of the benzene dimer was explored by on-the-fly molecular dynamics simulations based on the DFT-D method covering the dispersion energy. An all-electron DFT was performed at the BLYP/TZVP level. The parameters in the dispersion correction term were fitted to mimic the benchmark CCSD(T)/complete basis set limit potential energy curves for both the parallel-displaced (PD) and T-shaped (TS) structures of the dimer exactly. A dynamic description is important at temperatures above 10 K, where interconversion between the TS and PD structures is possible and a mixture of these two species exists. The higher the temperature, the more dominant the TS structure because of a favorable entropic contribution to the free energy. An analysis of the TS structures revealed that the symmetric $c_{2v}$ structure, a low-lying transition state, is practically not populated and that the tilted $c_s$ TS structure is prevalent. This finding is in perfect agreement with infrared spectroscopy.

## Introduction

A key role in biomacromolecules, organic and biomolecular crystals and nanomaterials is played by aromatic $\pi-\pi$ interactions,[1] a prototype of which is the benzene dimer. The potential energy surface (PES) of the benzene dimer contains two energy minima, the tilted T-shaped (TS) and parallel-displaced (PD) structures (see Figure 1 in our previous article[5]), which are separated by several transition structures. The relative stability between these minima has been the subject of dozens of experimental and theoretical papers, and its accurate determination requires an extremely high level of quantum mechanical description. Unfortunately, this information cannot be provided by experiments. The recent studies where the stabilization energies were determined at the CCSD(T)/complete basis set (CBS) level[2−6] or the DFT-SAPT/CBS level[7] have convincingly shown that both structures are practically isoenergetic. On the other hand, high temperature favors the T-shaped structure, as it is more flexible (with both the rotation and especially the tilting motions of the axial benzene) and therefore would be more stabilized than the stacked structure at nonzero temperatures owing to entropic effects.[4]

Yet another important fact is that the energy barriers separating the TS and PD energy minima, as well as the two TS minima, are very low, of the order of magnitude of 0.1 kcal/mol.[4,7,8] These extremely low barriers indicate that the concept of equilibrium structure is misleading and should be replaced by a dynamic average structure. This is even truer for nonzero temperatures, where in addition to the enthalpy term the entropy term becomes important. Since the vast majority of experiments are performed at nonzero temperatures, passing from the PES to the free energy surface (FES) is imperative. From the above-mentioned description of the PES of the benzene dimer, it is evident that a harmonic approach (like the rigid rotor−harmonic oscillator−ideal gas approximation) has several drawbacks. One possible solution is the application of molecular dynamics (MD) simulations. It is, however, evident that empirical potentials cannot describe the fine features of the PES and more accurate, quantum mechanical procedures are required. The choice of a suitable method is limited, since the dominant stabilization energy term in all benzene dimer structures is the London

* Corresponding author. E-mail: pavel.hobza@uochb.cas.cz.

10.1021/ct8002282 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

dispersion energy. It is now well-known that all standard DFT procedures fail to describe the dispersion energy[9,10] and only correlated methods like the MP2 or CCSD(T) can be utilized. These methods are expensive and cannot be applied in the on-the-fly MD simulations of such a large complex as the benzene dimer. However, the recently introduced DFT method augmented by the empirical dipsersion energy[11,12] (DFT-D) is a good approach to solving this problem.

Very recently, Pavone et al.[13] have studied the benzene dimer using atom-centered density matrix propagation ab initio molecular dynamics by means of the DFT-D method in the original Grimme's parametrization.[11] The authors presented 4 ps simulations at relatively high temperature (65 K) and found that the average structure converged rapidly to the T-shaped structure with the distance of the centers of mass being about 4.9 Å. However, several issues remain unresolved. First, Grimme's DFT-D underestimates the PD structure by about 0.6 kcal/mol as compared to the TS structure, whereas the most accurate QM methods (see above) show that the structures are practically isoenergetic. The "artificial" preference for the TS structure definitely affects the results of the presented simulations. Second, the system should be studied at temperatures ranging from very low to higher so that the temperature dependence of its behavior could be investigated. This dependence is also important for a comparison between the calculations and experiment, because the experimental conditions are not clearly defined. Finally, longer simulations, started from different minima, will bring more reliable statistics.

The goal of this paper is to investigate the PES and FES of the benzene dimer with the aim of describing the dynamic structure of the dimer at various temperatures. To this end, we carried out on-the-fly MD simulations based on the DFT-D procedure covering the dispersion energy, which exactly mimics the benchmark CCSD(T)/CBS for the PD and TS potential energy curves.

The determination of the dynamic structure of the benzene dimer is also topical in terms of an explanation of the nature of the C−H stretching mode shift observed in the infrared spectrum upon dimerization. It should be mentioned that the benzene dimer (T-shaped $c_{2v}$ structure) was the first system where we predicted[14] the improper blue-shifting hydrogen bond. While various examples of the blue-shifting bond were later detected experimentally,[15] the C−H$\cdots\pi$ blue-shifting bond in the benzene dimer was not confirmed. On the other hand, von Helden et al.[16] have recently found a small red shift (of $\sim$3 cm$^{-1}$) of the C−H stretch in the benzene dimer and shown that the two benzenes in the dimer are symmetrically inequivalent, which indicates a T-shape configuration. Very recently, we have suggested[17] that this may be explained by the fact that the $c_{2v}$ T-shaped structure, where we confirmed the existence of the blue shift, is a transition state. In an energy minimum, which is the tilted T-shaped ($c_s$ symmetry) structure, our anharmonic calculations suggest a red shift. However, the transition state is energetically very close to the minimum and may be accessible at higher temperatures.

## Methods

**DFT-D Calculations.** A description of the delicate balance between the conformations of benzene dimer requires an extremely accurate method. However, accurate methods tend to be computational expensive and thus unsuitable for MD simulations. The DFT-D method,[11,12] which uses transferable parameters applicable to a wide range of molecular complexes, can yield chemical accuracy but is still not accurate enough for a description of the benzene dimer. The next logical step is a parametrization of the DFT-D just on the benzene dimer, sacrificing the transferability of the parameters in order to reach the desired accuracy. Such parameters, derived in the preceding paper,[5] provide both the desired accuracy and efficiency and have been adopted for this study.

To maximize efficiency, this DFT-D procedure is based on the B-LYP[18,19] functional. Since a reasonably large basis set must be used to achieve the required precision, we selected the triple-$\zeta$ TZVP basis set.[20]

**On-the-Fly Molecular Dynamics.** On-the-fly molecular dynamics is also known as Born−Oppenheimer molecular dynamics, because it obeys the Born−Oppenheimer approximation. The electronic and nuclear motions are separated; the electronic structure is treated quantum mechanically, creating a potential for classical dynamics of the nuclei. In practice, this involves calculating the energy and gradients using the ab initio quantum mechanical (QM) method in each step of classical molecular dynamics.

We have developed our own molecular dynamics code that employs external programs to perform the QM calculations. The advantage of this design is that the MD algorithms and interfaces to external programs can be written in a high-level programming language, in this case Ruby, while the most time-consuming step, the QM calculation, is performed in software optimized for that task. As a result, the code is easy to extend and modify, yet very efficient. Modular design allows the use of various software packages for the actual calculation, but in this study, we consistently used Turbomole 5.9,[21] because it offers an extremely efficient implementation of the resolution of identity (RI) approximation with an optimized auxiliary basis set[22] within the DFT procedure. The interface to the QM packages includes the implementation of the empirical dispersion calculation within the DFT-D scheme and is thus independent of the actual code used for calculation.

**Energy Conservation in On-the-Fly Ab Initio MD.** The first and most obvious parameter affecting energy conservation is the time step. We tested several values in simulations at constant energy. In this study, we used a 1 fs step, which had proven to be short enough to conserve total energy well while maximizing the accessible simulation time scale.

On-the-fly MD involve one more problem with energy conservation unknown in molecular mechanics—residual gradients caused by an imperfect convergence of the SCF procedure in the QM calculation. Some error due to the finite convergence limit always occurs, and Pulay and Fogarasi[23]

have shown that this error accumulates during the simulation if molecular orbitals from the previous step are used as an initial guess for the calculation. This systematic error leads to serious leaks of kinetic energy in the time scale of picoseconds. To avoid this problem, we started the QM calculation from scratch in each step. In this case, the convergence error becomes stochastic and is canceled out during the simulation. The SCF convergence limit is set to the default value of $10^{-6}$ Hartree. However, such a calculation takes roughly twice as much time as starting from the previous step, but this is the best solution possible within our implementation.

With the described setup, we were able to run 20 ps constant energy simulations of the benzene dimer with a perfect conservation of total energy.

**Selection of a Thermostat Algorithm.** Since the time scale accessible to a molecular dynamics simulation at this level is relatively short, extra care must be taken to ensure proper sampling of conformational space. We have found that the selection of the thermostat algorithm is the most important choice affecting the results, especially in the gas phase. When a thermostat based on a global scaling of velocities, such as the commonly used Nosé−Hoover algorithm, is utilized, the only path of the redistribution of the kinetic energy in the molecule is an internal energy flow, which is slow. As a result, the conversion between the accessible conformations does not occur or takes a very long time. This effect has not been observed in condensed phase simulations, where the solvent facilitates the energy transfer.

To overcome this problem, an algorithm that does not conserve the direction of momentum should be applied. We use the Andersen thermostat, which simulates random collisions with a thermal bath. During the simulation, the velocity of a randomly selected atom is newly generated from a Maxwell distribution for a desired temperature with a selected average collision frequency. As a result, the average temperature as well as the kinetic energy distribution is conserved, but the momentum direction is randomly modified several times during the simulation. This makes it possible for the system to escape from the local minimum and thus leads to more efficient sampling.

**Simulation Protocol.** All the simulations presented here were 20 ps long, with a time step of 1 fs. The Verlet propagation algorithm was used in conjunction with the Andersen thermostat.[24] The mean collision frequency of the thermostat was set to 2.5 $ps^{-1}$.

In this work, we explored temperatures ranging from 10 to 100 K in 10 K increments. This was necessary for proper comparison with experiment. The temperature in the molecular clusters after expansion was definitely very low (∼10 K). However, in almost all cases, the pre-expansion (nozzle) temperatures were much higher, which enabled sampling of a greater variety of species. The temperature gradient during the expansion was large enough to freeze the structures populated only at higher temperatures instantaneously.

At each temperature, eight MD simulations were run. Since there are two main minima corresponding to the TS and PD structures, we started half of the simulations from one minimum and half-from the other. At low temperatures, we explored only two structures, whereas at temperatures above the interconversion barrier, this allowed us to check how the trajectory depended on the initial structure. At 80 K, we doubled the number of simulations to check the reliability of the averaging.

The trajectories in one set differed only in the random initial velocities, generated from a Maxwell distribution for the given temperature.

**Probability Distribution of Structural Parameters.** Raw trajectories were processed to obtain a probability distribution of the structural parameters at a given temperature. First of all, we had to ensure that the collected results were not affected by the initial structure. To assess how long a time scale it would take to lose this correlation, we calculated the autocorrelation function c of the Cartesian displacement coordinates $q$

$$q_i = x_i - \bar{x}_i \tag{1}$$

and variance $\sigma^2$ of the coordinates in the simulation with $N$ frames

$$\sigma^2 = \frac{1}{N}\sum_{t=1}^{N} (x(t) - \bar{x}) \cdot (x(t) - \bar{x}) \tag{2}$$

where $x(t)$ denoted the geometry in frame $t$ as

$$c(t) = \frac{\langle (x(t) - \bar{x}) \cdot (x(0) - \bar{x}) \rangle}{\langle \sigma^2 \rangle} \tag{3}$$

where angle brackets denoted the average over multiple trajectories differing in their initial conditions.

To obtain a result independent of the initial geometry, we could take only a part of the trajectory beyond the point where the autocorrelation approaches zero. In practice, we always discarded the first half (10 ps) of each trajectory after we determined that the autocorrelation decayed in approximately 5 ps. We did so because our autocorrelation functions were not perfect as they had been averaged over a small set of trajectories.

In the remaining part of the trajectory, the desired parameter was measured in each frame. A normalized histogram (probability distribution) of these values was built and averaged over the trajectories in the set.

## Results and Discussion

A total of 88 trajectories were calculated, which took a total CPU time of 3.4 CPU years. All the trajectories were processed using the above-described protocol to reduce the data to probability distributions (histograms) of the structural parameters for each combination of temperature and initial structure.

First of all, we looked at populations of the T-shaped and parallel-displaced minima. The histogram of the angle between the benzene plane rings (α, see Figure 1), a coordinate best distinguishing these two structures, is plotted in Figure 2. Degrees of 0 and 90 correspond to the PD and TS structures, respectively. In this coordinate, the peak at 90° covers all the possible T-shaped structures, which may differ in the tilt angle (see the discussion below). Separate analyses were conducted for simulations starting from the

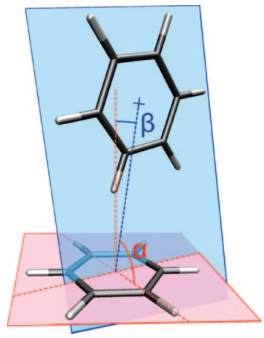



***Figure 1.*** Definition of the angle between the ring planes α and the tilt angle $\beta$.

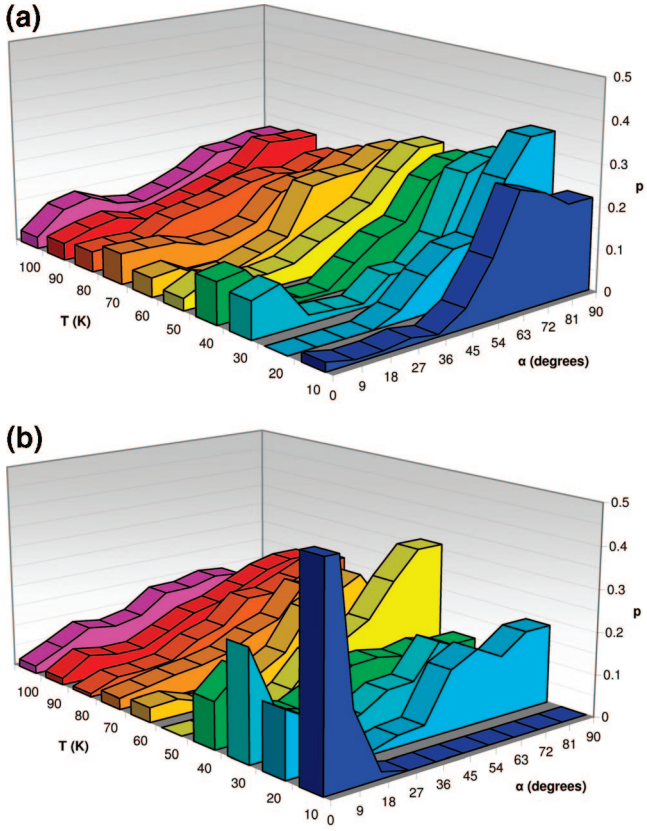



***Figure 2.*** Probability distribution (histogram) of angle α between the benzene rings in the dimer, plotted for various temperatures. Degrees of 0 and 90 correspond to the PD and T-shaped structures, respectively. The simulations were started from the TS (a) and PD (b) minima.

T-shaped structure (a) and the parallel-displaced structure (b). Due to the short length of the trajectories, the histograms are not always smooth and there are some outlying points, yet the trends are well resolved.

Before we begin an analysis of the simulations, we should look at the benzene dimer at 0 K. The interaction energy suggests that the tilted T-shape is slightly more stable (by 0.08 kcal/mol in the DFT-D potential used in our simulations) or practically isoenergetic, as discussed above. When we add the zero-point vibration energy (ZPVE) and compare the resulting enthalpy at 0 K, these two structures become isoenergetic (see ref 4 for further details) or the PD structure is slightly favored (by 0.15 kcal/mol in our potential). On the other hand, experiment at temperatures close to 0 K (achieved by jet cooling) records the T-shaped structure of benzene dimer. The question remains whether the vibrational temperature is really so low (see above). An accurate theoretical description of the benzene dimer at 0 K is yet to be achieved, mainly through a more accurate calculation of the lowest vibrational modes.

It must be noted that the harmonic calculation of the lowest vibrational frequencies is only a crude approximation, because these vibrations are anharmonic or nonharmonic (the anharmonic treatment is still an approximation, although of a higher order; the nonharmonic calculation would pose no limitations on the potential form). These limitations are absent in molecular dynamics: True nonharmonic potential is explored in the simulation. We also attempted to go beyond the harmonic approximation in the static description by using perturbation theory[25] as implemented in Gaussian 03,[26] but these calculations failed.

At the lowest simulated temperature of 10 K, the barrier between the TS and PD structure cannot be overcome, and the dimer stays in the minimum from which the simulation started. Above this temperature, interconversion between TS and PD is possible. There is an interesting region between 20 and 40 K, where a well-resolved peak at 0° corresponding to the PD structure is observed even in simulations starting from the T-shape. Here, both structures can coexist, although the T-shaped minimum is more populated.

At higher tempereatures (from 50 K and above), however, the results are unambiguous and the T-shaped structures are clearly favored. The question arises what the driving force for this preference is if both structures are practically isoenergetic. The answer is clear—the entropy is what turns the balance toward the T-shaped structure. Although the stacked structures are usually entropically favored when compared to the hydrogen-bonded ones, e.g. in DNA base pairs, the CH$\cdots\pi$ interaction in the benzene dimer is not rigid at all. The entropy gain as compared to the PD structure comes mainly from two sources: an almost free rotation of the horizontal (in the T letter) benzene and a very soft tilting vibrations of the vertical one (causing a very broad distribution of the T-shaped structure in the plane angle histograms even at low temperatures). Both benzene rings do rotate in both structures, but particularly this rotation is observed at the lowest temperature. The stacked arrangement is more rigid because the whole rings are involved in the interaction. As a measure of this rigidity, we have calculated harmonic frequencies of the six intermolecular vibrational modes. For the TS structure, the average value is 80 cm$^{-1}$, while in PD, the average is 87 cm$^{-1}$. We believe that this difference will be even more pronounced in an anharmonic calculation.

The entropy can be estimated on the basis of the MD simulations performed. Although the length of the trajectories is not sufficient for an accurate analysis, it is enough to show the trends and provide semiquantitative results. For this analysis, the border between the PD and TS structures should be defined. One possibility is to divide the range into halves and set the threshold for further analysis to 45°. The other possibility is to use the minimum from the plane angle histograms at about 30° (cf. Figure 2). Both values lead to similar results; the calculation presented below is based on

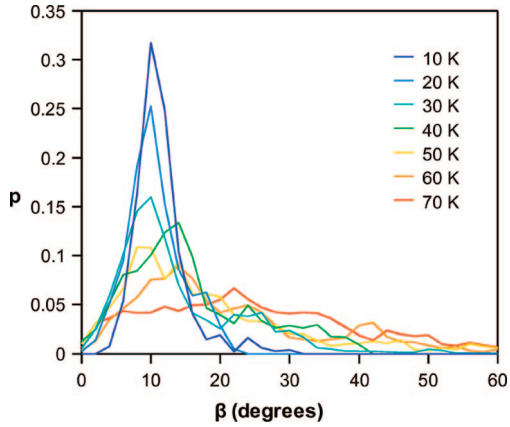



***Figure 3.*** Population ($p$) of tilt angle $\beta$ in simulations of the T-shaped benzene dimer at various temperatures.

the 45° threshold. At each temperature, we calculated the equilibrium constant K as a ratio of the TS and PD structures. From the equilibrium constant, we can readily obtain the free energy difference

$$\Delta G = -RT \ln K \quad (4)$$

From a linear regression of the temperature dependence of $\Delta G$, we can extrapolate to 0 K to obtain $\Delta H^0$. We found the TS structure disfavored by 0.035 ± 0.05 kcal/mol at 0 K. From the slope of the fitted line, we obtained an entropy difference $\Delta S = 0.004 \pm 0.001$ kcal/(mol K). At 100 K, the entropic stabilization of the TS structure thus amounted to −0.4 kcal/mol.

This result is extremely important for the interpretation of the experiments, because it shows that the T-shaped conformation is more populated even at low temperatures and prevalent at high temperatures (which is important even in experiments at low temperature, as discussed above).

However, the measure discussed above does not distinguish between the tilted T-shaped global minimum ($c_s$) and the symmetrical transition state ($c_{2v}$). To solve this issue, we measured tilt angle $\beta$ (0° for the symmetrical structure, 7.4° for the tilted global minimum; cf. Figure 1) in the trajectories starting from the TS structure, where it could be easily defined. The results, summarized in Figure 3, show that there is only a negligible population of the symmetrical structure at higher temperatures and none at 10 K. Passing from 10 K to higher temperatures, the average tilt angle gradually increases above the equilibrium value of 7.4°. At temperatures above 30 K, the trajectories sample both minima, and angle $\beta$ no longer describes only the TS structure.

This finding makes it possible for us to interpret the experimental results of von Helden et al. (see above). Evidently, the small red shift which they found should be assigned to the prevalent tilted T-shaped structure. This is in agreement with our previous calculations,[17] where we predicted a red shift for this structure; a blue shift was assigned to the $c_{2v}$ structure, which is only negligibly populated in our simulations.

Another parameter studied was the distance of the centers of masses of the benzene monomers. It distinguishes perfectly between the parallel-displaced (3.90 Å in the minimum) and T-shaped (4.90 Å in the minimum) structures. The value for the T-shaped structure, calculated by our DFT-D method, agrees well with the experimentally measured[27] distance of 4.9 ± 0.01 Å, determined from the rotational spectrum.

***Table 1.*** Average Distance of the Benzene Centers of Mass in the Parallel-Displaced (PD) and T-Shaped (TS) Dimers from MD Simulations at Different Temperatures

| | distance (Å) | |
|---|---|---|
| T(K) | PD | TS |
| 10 | 3.95 | 4.87 |
| 20 | 3.98 | 4.88 |
| 30 | 4.00 | 4.87 |
| 40 | 4.03 | 4.89 |
| 50 | 4.07 | 4.93 |
| 60 | *a* | 4.88 |
| 70 | 4.12 | 4.98 |
| 80 | 4.10 | 4.99 |
| 90 | 4.13 | 5.01 |
| 100 | 4.16 | 4.99 |

[a] No parallel-displaced structure was observed in the simulation.

From molecular dynamics, we can extract the average distance for each structure at different temperatures. The results, listed in Table 1, show an increase of the distance with temperature. This proves the anharmonic nature of this intermolecular mode, and the rate of this increase (higher in the PD structure) could serve as a measure of anharmonicity.

## Conclusions

• The DFT-D method offers a very good accuracy-to-computational-cost ratio. The parameters customized precisely for the benzene dimer made it possible for us to run on-the-fly molecular dynamics simulations with a precision of coupled-cluster calculations.

• The static description of the benzene dimer at a temperature of 0 K based on an accurate energy calculation and a harmonic calculation of the vibrational frequencies suggests that the parallel-displaced structure is slightly more stable, whereas experiment detects the T-shaped structure. This issue is yet to be resolved, namely by an accurate anharmonic calculation of the vibrational frequencies and the zero-point vibrational energy.

• The dynamic description becomes important at temperatures above 10 K, where interconversion between the TS and PD structures becomes accessible. At low temperatures, a mixture of these two configurations exists, but the T-shaped structure becomes dominant as temperature increases, because it is favored by the entropic contribution to the free energy. This finding explains the experimental studies which detected only the TS structure provided that the vibrational temperature of the cluster was above 0 K, which is probable in jet cooling experiments.

• The T-shaped structure is tilted and the fully symmetric $c_{2v}$ T-shaped transition state remains unpopulated even at high temperatures. This is in full agreement with infrared spectroscopy and the calculations that assigned the spectrum to the tilted TS.

• The intramolecular distance measured in our calculations is in perfect agreement with the experimental value; its

increase with temperature provides evidence of the anharmonicity of the intermolecular potential.

**Acknowledgment.** This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (Grant No. LC512) and by the Grant Agency of the Academy of Sciences of the Czech Republic (A400550510); it was also part of the research project No. Z4 055 0506. The support of Praemium Academiae of the Academy of Sciences of the Czech Republic, awarded to P.H. in 2007, is also acknowledged.

## References

(1) Müller-Dethlefs, K.; Hobza, P. Noncovalent interactions: A challenge for experiment and theory. *Chem. Rev.* **2000**, *100* (1), 143–167.

(2) Distasio, R. A.; von Helden, G.; Steele, R. P.; Head-Gordon, M. On the T-shaped structures of the benzene dimer. *Chem. Phys. Lett.* **2007**, *437* (4−6), 277–283.

(3) Janowski, T.; Pulay, P. High accuracy benchmark calculations on the benzene dimer potential energy surface. *Chem. Phys. Lett.* **2007**, *447* (1−3), 27–32.

(4) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. Understanding of assembly phenomena by aromatic-aromatic interactions: Benzene dimer and the substituted systems. *J. Phys. Chem. A* **2007**, *111* (18), 3446–3457.

(5) Pitoňák, M.; Neograd y, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. Benzene dimer: High-level wavefunction and density functional theory calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1829–1834.

(6) Sinnokrot, M. O.; Sherrill, C. D. Substituent effects in pi-pi interactions: Sandwich and T-shaped configurations. *J. Am. Chem. Soc.* **2004**, *126* (24), 7690–7697.

(7) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Potential energy surface for the benzene dimer and perturbational analysis of $\pi-\pi$ interactions. *J. Phys. Chem. A* **2006**, *110* (34), 10345–10354.

(8) Špirko, V.; Engvist, O.; Soldan, P.; Selzle, H. L.; Schlag, E. W.; Hobza, P. Structure and vibrational dynamics of the benzene dimer. *J. Chem. Phys.* **1999**, *111* (2), 572–582.

(9) Hobza, P.; Šponer, J.; Reschel, T. Density-Functional Theory and Molecular Clusters. *J. Comput. Chem.* **1995**, *16* (11), 1315–1325.

(10) Kristyan, S.; Pulay, P. Can (Semi)Local Density-Functional Theory Account for the London Dispersion Forces. *Chem. Phys. Lett.* **1994**, *229* (3), 175–180.

(11) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25* (12), 1463–1473.

(12) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *J. Comput. Chem.* **2007**, *28* (2), 555–569.

(13) Pavone, M.; Rega, N.; Barone, V. Implementation and validation of DFT-D for molecular vibrations and dynamics: The benzene dimer as a case study. *Chem. Phys. Lett.* **2008**, *452* (4−6), 333–339.

(14) Hobza, P.; Špirko, V.; Selzle, H. L.; Schlag, E. W. Anti-hydrogen bond in the benzene dimer and other carbon proton donor complexes. *J. Phys. Chem. A* **1998**, *102* (15), 2501–2504.

(15) Hobza, P.; Havlas, Z. Blue-shifting hydrogen bonds. *Chem. Rev.* **2000**, *100* (11), 4253–4264.

(16) Erlekam, U.; Frankowski, M.; Meijer, G.; von Helden, G. An experimental value for the B-1u C-H stretch mode in benzene. *J. Chem. Phys.* **2006**, *124* (17), 171101–171104.

(17) Wang, W. Z.; Pitoňák, M.; Hobza, P. C-H stretching vibrational shift of benzene dimer: Consistency of experiment and calculation. *ChemPhysChem* **2007**, *8* (14), 2107–2111.

(18) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

(19) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula Into a Functional of the Electron-Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

(20) Schafer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829–5835.

(21) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165–169.

(22) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. *Theor. Chem. Acc.* **1997**, *97* (1−4), 119–124.

(23) Pulay, P.; Fogarasi, G. Fock matrix dynamics. *Chem. Phys. Lett.* **2004**, *386* (4−6), 272–278.

(24) Andersen, H. C. Molecular-Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393.

(25) Barone, V. Anharmonic vibrational properties by a fully automated second-order perturbative approach. *J. Chem. Phys.* **2005**, *122*, 1.

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

(27) Arunan, E.; Gutowsky, H. S. The Rotational Spectrum, Structure and Dynamics of a Benzene Dimer. *J. Chem. Phys.* **1993**, *98* (5), 4294–4296.

CT8002282

JCTC Journal of Chemical Theory and Computation

# Chiral Aromaticities. A Topological Exploration of Möbius Homoaromaticity

Charlotte S. M. Allan and Henry S. Rzepa*

*Department of Chemistry, Imperial College London, South Kensington Campus, London SW7 2AZ, U.K.*

Received May 27, 2008

Ⓦ **This paper contains enhanced objects available on the Internet at http://pubs.acs.org/JCTC.**

**Abstract:** A series of $C_2$-symmetric homoderivatives of the cyclo $C_9H_9^+$ cation first identified by Schleyer as Möbius aromatic are shown to themselves sustain Möbius 4n-$\pi$-electron homoaromaticity. Analogous double-twist Möbius bis-homoaromatics follow a 4n+2 electron rule. AIM (atoms-in-molecules) and ELF (electron localization function) analysis of the electron topology in the region of the homobond of these systems reveals that the presence of a AIM bond-critical point in this region is not mandatory, it being unstable to subtle variations in the local electron density induced by local or remote substituents, and which can in turn induce self-annihilation or creation of a pair of bond and ring critical points. The same substituent-induced annihilation/creation of such a BCP/RCP pair can also be observed in the nonclassical norbornyl cation. We suggest that the ELF and $ELF_\pi$ thresholds for any basin found in the homoregion are better indicators of the delocalized nature of the homoaromatic interaction and the aromaticity of the system.

## 1. Introduction

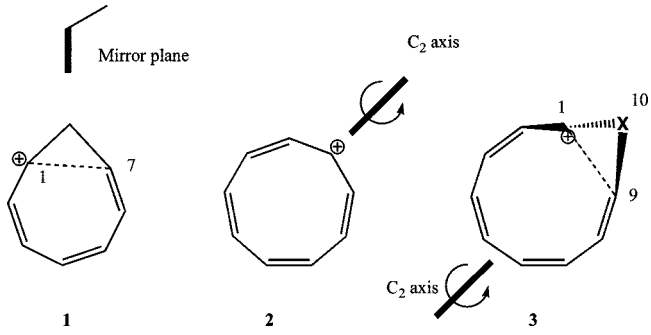
Homoaromaticity is a term introduced by Winstein[1] in 1959 to describe systems in which the $\sigma$ framework of a cyclic conjugated and planar aromatic (4n+2) $\pi$-electron framework is interrupted with one or more bridging groups (predominantly but not necessarily manifested by $CH_2$). This results in through space rather than through bond overlap of the $p_\pi$-$p_\pi$ framework. Since then a large body of work has concluded that the phenomenon manifests best in cationic systems, of which the homotropylium ion **1** is held as the archetypal example.[2] Homoaromaticity as described by Winstein has thus far been exclusively interpreted in terms of achiral benzenoid models in which the $p_\pi$-$p_\pi$ overlaps occur with preservation of an (idealized) plane of symmetry, the so-called Hückel aromaticity model. Yet there is another model, a chiral (dissymmetric) one characterized as having (idealized) axes of symmetry only, and which has become known as the 4n-electron Möbius aromaticity model.[3] This distinction was first clearly introduced as a convenient selection rule for the transition states of organic pericyclic reactions[4] in 1966. Only in the past decade however has it has been extended to characterized examples of stable organic Möbius-aromatic molecules.[5] Most recently further diversification to a variety of organometallic, metallacyclic, and inorganic substances has been reported.[6] The first (and still one of the best) examples of cyclic Möbius $\pi$-conjugation to be identified is the $C_9H_9^+$ cation, shown by Schleyer[7] to be an 4n eight $\pi$-electron delocalized and chiral aromatic cycle **2** bearing a $C_2$ axis of symmetry only. Here we extend the diversity of Möbius behavior to that of Möbius homoaromaticity, using as the starting point a theoretical exploration of systems derived from **2**.

## 2. Results and Discussion

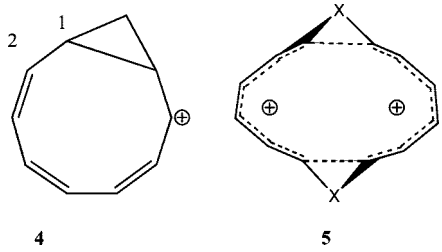
**2.1. The Möbius Homoaromatic 3, X=CH$_2$.** Combining the characteristics of **1** and **2** results in the $C_2$-symmetric 8-$\pi$-electron **3** (X=$CH_2$), in which the $\pi$ system undergoes a half-twist in the cycle. Any homoaromaticity would be defined by the degree of conjugation sustained across the C1−C9 bond in this species and whether any resulting ring current in the system is strongly diatropic. The latter is most simply quantified by the NICS index introduced by Schley-

* Corresponding author e-mail: rzepa@ic.ac.uk.

10.1021/ct8001915 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008




er.[8] Preliminary results for the prototypic system **3**, X=$CH_2$ were reported elsewhere,[9] B3LYP/6−31G(d,p) calculations indicating that the C1−C9 bond was surprisingly short ( Web-enhanced Table 1), but that this value was *also* associated with a low degree of bond length alternation (i.e., geometric delocalization) in the ring defined by the $sp^2$-hybrid carbon atoms. This alternation can be simply approximated by specifying the difference between the shortest and longest bonds in this cycle ($\Delta_r$) and has the value of 0.035 Å in this instance. If in fact the structure was better represented by the bicyclic isomer **4** (with a formal $\sigma$-single bond between C1−C9), the interrupted delocalization would only extend between C2−C8, and the C1−C2 bond would also be close to a single bond in length (it actually has a predicted value of 1.419 Å, close to the typical aromatic value in benzene). This aspect can be tested further by calculating the anion of **3**, X=$CH_2$, which as a 4n+2 10-electron half-twist system would yield a Möbius homoantiaromatic. The molecule avoids this (high energy) solution by instead adopting the anionic form of **4**, for which the C1−C9 and C1−C2 bonds are indeed both long (1.502 and 1.480 Å, respectively, $\Delta_r$ 0.113Å) and sustaining eight acyclic rather than ten cyclically conjugated $\pi$-electrons. Winstein has previously noted analogous behavior for metallocomplexes of **1**.[10] Thus while **1** as coordinated by $Cr(CO)_3$ (a 6$\pi$-acceptor) is genuinely homoaromatic, when the ligand is replaced by $Fe(CO)_3$ (a 4$\pi$-acceptor), the complex instead adopts the bicyclic form with a formal single bond between C1−C7.




To characterize the nature of the C1−C9 bond in **3** (X=$CH_2$), we used Bader's critical point analysis[11] of the electron density (AIM) and the related ELF[12] (electron localization function). AIM (Atoms-in-molecules) involves analysis of the rate of curvature (Laplacian) of the electron density $\rho$(r) in terms of four types of so-called critical points, at each of which the derivative of $\rho$(r) is zero. These four are nuclear critical points (located at the nuclei), bond critical points (BCP) located between (normally pairs of) nuclei, ring critical points (RCP, defining a ring of nuclei), and cage critical points (CCPs). A topological relationship (the Poincaré-Hopf rule) between the numbers of each type of critical point states that NCP - BCP + RCP - CCP = 1. This method had previously[13] been applied to **1**, with the surprising result that no BCP can be identified in **1** along the path connecting C1 to C7. In recognition of this feature, **1** has become known as a no-bond homoaromatic species.[13] It remains contentious whether the topological interpretation of the electron density provided by this analysis necessarily relates to the best *chemical* description of the bonding. In any event, **3** also exhibits no bond critical point in the C1−C9 region (Figure 1), although the predicted separation of these nuclei is in fact much shorter than that for **1** (Web-enhanced Table 1). The AIM analysis does however provide precisely one RCP for **3** (X=$CH_2$), and its coordinates provide a convenient location for measuring the magnetic properties of this system via the NICS (nucleus independent chemical shift).[8] Thus the NICS(rcp) of **3** has the value of −11.3 ppm (in comparison, benzene has a value at the equivalent ring centroid of ∼−10, **1** sustains a value of −11.5 and **2** of −10.9 ppm). By these various measures, **3** (X=$CH_2$) is clearly aromatic and more specifically homoaromatic.

It has recently been argued that a better chemical interpretation is provided not by the topology of the full electron density but by a related measure known as the electron localization function (ELF(r)).[15] This function provides information relating to localized electron pairs and therefore gives a direct insight into chemical bonding. In the density functional approach we are using here, ELF is calculated using the Pauli kinetic energy density. A consequence of the Pauli principle is that the probability of finding two electrons of opposite spin in the same region of space is increased, which leads to greater Coulomb repulsion between these two electrons and results in a larger kinetic energy (than if the Pauli principle had been ignored). It is this excess electronic kinetic energy value, relative to a homogeneous electron gas (of the same density), that defines the DFT-based ELF, which has been formulated to take values between 0 and 1. The greater the excess kinetic energy (and therefore the greater the Pauli repulsion and concomitant likelihood of finding pairs of electrons of opposite spin), the closer the ELF function is to 1. As with AIM, critical points can be identified in the properties of this function, but in fact a more useful analysis is the localization domain reduction tree (LDRT), which is used to identify so-called bifurcation thresholds for basins in the ELF topology.[15a,b] As ELF(r) is increased from 0 toward 1.0, the thresholds at which the ELF basin in the region of any putative bond first appears (bifurcates) and then vanishes are useful indications of the nature of the chemical bonding there. Thus the all-electron valence basins for most conventional bonds vanish to a point at ELF(r) values of ∼0.94−0.95, whereas weaker interactions such as hydrogen bonds, agostic interactions, or $\pi$-stacking may not even sustain the formation of discrete basins. The basins themselves can be one-centered (monosynaptic; lone pairs), two-centered (disynaptic), or three-centered (trisynaptic). Appropriate integration over the volume of any basin yields the total number of electrons associated with each basin.

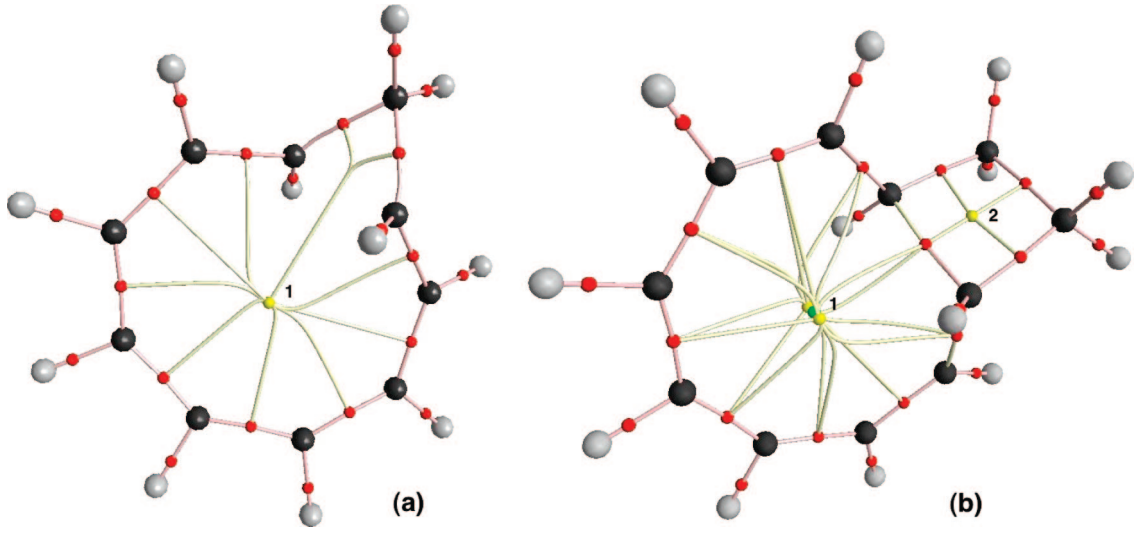



***Figure 1.*** Critical points for (a) **3**, X=$CH_2$ and (b) **3**, X=$CH_2$ $CH_2$. Bond critical points are shown in red, ring critical points in yellow, and cage critical points in green. The $\rho$(r) values (in e.Å$^{-3}$, B3LYP/6−31G(d,p)), at RCP points (yellow) 1 are 0.004, and 0.072 at point 2. In (b), the helicity of the system causes the RCP to split into two such points, connected by a CCP (green). Such behavior has been previously noted.[14]

When this procedure is applied to **3**, X=$CH_2$, a disynaptic all-electron basin is identified in the region of the C1−C9 bond and is designated V(C1,C9). It has a ELF(r) 0.675 for its bifurcation point, 0.776 at its vanishing point, and a value of 0.79 electrons for the basin integration (Web-enhanced Table 1). This relatively low excess value for the Pauli repulsion energy does correspond to the chemical interpretation that the C1−C9 "bond" is significantly more delocalized than a conventional single bond type. The 0.79e basin population corresponds nicely to the concept of a $\pi$-bond shorn of its underlying 2-electron $\sigma$-framework (i.e., a homoaromatic bond); the V(CC) basin integrations for the other C−C bonds in the ring indeed do range from 2.4 to 2.9e, corresponding to more conventional two-center-three-electron aromatic bonds.

It is also possible to perform the ELF analysis by specifying a subset of the occupied molecular orbitals, such as the $\pi$-manifold.[16] Four occupied MOs for **3**, X=$CH_2$ can be identified as $\sim\pi$ (Figure 2) although mixing from the C−H $\sigma$-manifolds is also clearly apparent (*e.g.* Figure 2d). Simple Huckel MO theory predicts that $\pi$-orbitals in a Möbius cycle will occur as degenerate pairs,[3] although in practice this degeneracy is always broken.[17] The most stable $\pi$-MO pair (orbitals 31, 32, Figure 2) each has one node which occurs at one of the bisection points of the cycle by the $C_2$ axis of symmetry, and the less stable of this pair (albeit by only 0.1 eV, Figure 2c) has this node in the C1−C9 homobond region, again coincident with the $C_2$ axis. This type of behavior is also manifested by the parent ion **2**.[17]

ELF(r)$_\pi$ obtained using just this orbital subset exhibits bifurcation values for most of the $\pi$-like synaptic basins over the range 0.80−0.88, compared to the characteristically aromatic value for V(C,C)$_\pi$ of 0.91 reported for benzene.[16] There are differences however in behavior compared to a planar aromatic such as benzene. For example, whereas the benzene ELF(r)$_\pi$ bifurcation into two equal but separate disynaptic $\pi$-valence basins (one above, the other below the plane of the ring) must occur at the same threshold, the process for a helical Möbius system involves no such plane of symmetry and the two opposing lobes of the $\pi$-system may separate at different thresholds. This difference is illustrated by the ELF(r)$_\pi$ isosurface at 0.47 for **3**, X=$CH_2$ (Web-enhanced Table 1), which represents the bifurcation of the "inside" or endocyclic lobe of the C1−C9 $\pi$-basin away from its neighbors, but leaving a 5-center basin representing much of the rest of the original cyclic $\pi$-density still unseparated (Figure 4). The previously reported ELF(r)$_\pi$-based aromaticity scale[16] interprets a value of 0.47 as verging on the antiaromatic, but the low bifurcation value for this particular ELF basin may instead be characteristic of either homoaromaticity or of Möbius topology or both. Thus the corresponding values for **2** (0.33 for the initial $\pi$-bifurcation and 0.95 for its completion) are very similar to those for **3**, X=$CH_2$. What is clear is that while a simple Möbius aromaticity scale based on ELF(r)$_\pi$ values is unlikely to be accurately quantitative, the ELF-inferred homoaromaticity of **3**, X=$CH_2$ certainly approximates to that of benzene in most regards.

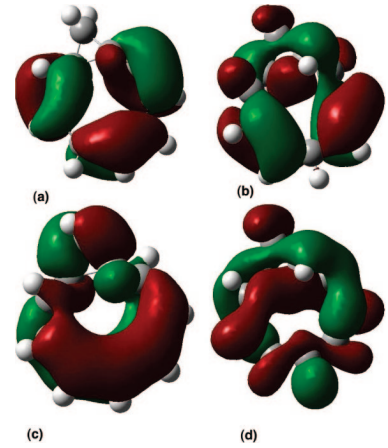



***Figure 2.*** $\pi$-like molecular orbitals for **3**, X=$CH_2$ for (a), orbital 35, (b), orbital 34, (c) orbital 32, and (d) orbital 31, contoured at a threshold of 0.02 au at the B3LYP/6−31G(d,p) level.

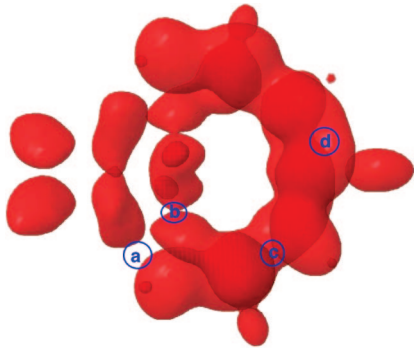



***Figure 3.*** ELF$_\pi$=0.495 isosurface for **3**, X=CH$_2$, revealing that bifurcation of the $\pi$-surface at point (a) has already passed, whereas those at points (b) and (c) are on the cusp, and that at e.g. point (d) and others are still some way off. The isosurface prior to bifurfurcation has the topology of a one component torus knot.[17]

The preceding ELF and AIM analyses have revealed a clear divergence for **3**. Whereas no AIM C1−C9 bond critical point exists, ELF shows a clear disynaptic valence basin in this region, occupied by ∼1 $\pi$-electron. This contrasts with the AIM and ELF analyses (Web-enhanced Table 1) of the homotropylium cation **1**, the former sustaining no C1−C7 bond critical point, and the latter concurring with absence of a C1−C7 disynaptic basin. Therefore while **1** is indeed well described as a no-bond homoaromatic system, **3** arguably should not be so described.

Another measure of a Möbius electronic system is the properties of its $\pi$-AO ribbon as defined by the cyclic array of the p-atomic orbital basis and a topological property of that ribbon known as the linking number L$_k$. An important theorem introduced by White, Fuller, and Cãlugãreanu[18] defines a decomposition of L$_k$ into noninteger total twist (T$_w$) and writhe (W$_r$) components, according to

$$L_k = T_w + W_r$$

The two (noninteger) components of L$_k$ are obtained by integration of the appropriate functions of the ribbon coordinates, with T$_w$ being the integral of all local torsions $\gamma$ around the center line of the basis set ribbon and W$_r$ (known as the writhe of the system) being a nonlocal property obtained by a double-integration.[19] The writhe of the ring describes the extent to which the center line of the basis set ribbon projects from 2D into 3D space. Expressed in units of $\pi$, a value of L$_k$ =1 corresponds to the conventional description of a single half-twist Möbius system, while larger values are often referred to as the corresponding higher-order Möbius systems. Using the previously described protocol,[19] the values for **3**, X=CH$_2$ are computed as L$_k$ = 1$\pi$, T$_w$=1.13, and W$_r$ = −0.13, this being the first homoaromatic system analyzed in this manner. The degree by which W$_r$ modifies the value of T$_w$ (they are normally but not necessarily of the same sign) can be most simply interpreted in terms of how much the (unfavorable) reduction in p$_\pi$-p$_\pi$ AO overlaps brought about by T$_w$ torsion can instead be converted into (presumably the more favorable) bending deformations which are a feature of writhe.[19] In this example, little such conversion has taken place; the system being almost a pure half-twist Möbius molecule, with little projection of the torsions into the writhe of 3D space. This might be contrasted with the values for **2** itself, which were reported[19] as L$_k$ = 1$\pi$, T$_w$=0.73, and W$_r$ = 0.27$\pi$, indicating some relief of (overlap reducing) p$_\pi$-p$_\pi$ torsions.

**2.2. Other Möbius Homoaromatics Based on 3.** With these methodologies established for characterizing the species, we next explored the results of varying the nature of the bridging group X. For the larger bridge X=CH$_2$−CH$_2$, similar AIM results were obtained in all but one regard. The C1−C9 bond length was slightly longer, $\Delta_r$ was somewhat larger, but the NICS(rcp) was still characteristic of an aromatic system. However, the system now **did** exhibit (Figure 1b) a BCP in the C1−C9 region, despite the nuclei being further apart! The ELF(r) V(C1,C9) basin disappearance threshold of 0.879 and a basin integration of 1.17e also indicated somewhat more localized bonding, tending toward **4**.

Our task now became one of exploring whether chemical characteristics for predicting whether a BCP might be expected in any given bonding region can be defined. This was achieved by systematically varying the electron demand of X, *via* both the central atoms, and their substituents. Thus **3**, X=C(SiH$_3$)$_2$ increases the electron donation of the substituent, and the critical point analysis now changes quantitatively, in revealing two additional critical points compared to X=CH$_2$, comprising one BCP and one RCP. It

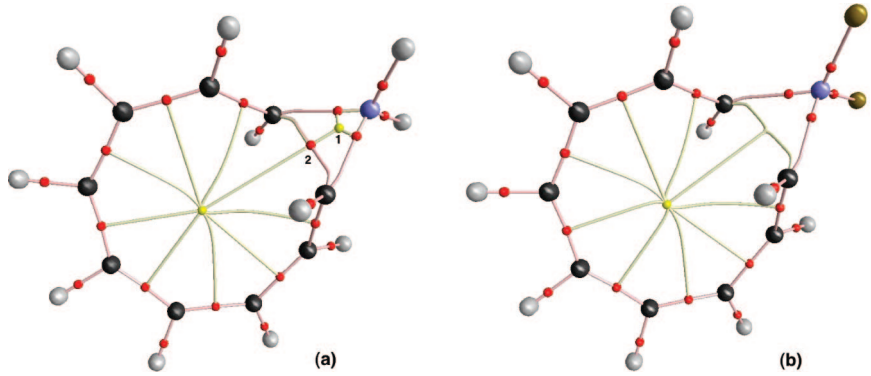



***Figure 4.*** Critical points for (a) **3**, X=BH$_2$ and (b) **3**, X=BF$_2$. The $\rho$(r) values (in e.Å$^{-3}$, B3LYP/6−31G(d,p)), at RCP point (yellow) 1 and BCP point 2 (red) are 0.134 and 0.140, respectively. In (b), these two points have self-annihilated as a result of replacing H by F.

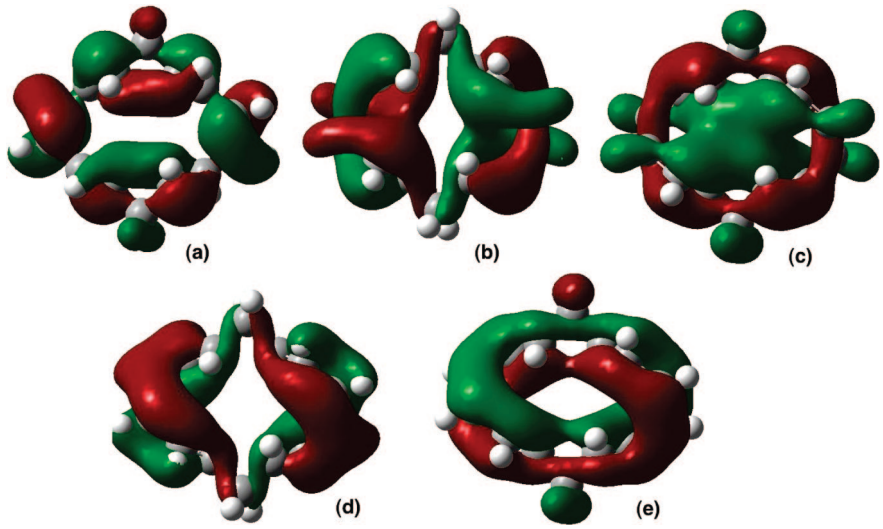



***Figure 5.*** $\pi$-like molecular orbitals for **5**, X=$CH_2$ for (a), orbital 49, (b), 48, (c) 47, (d) 46, and (e) 45, contoured at a threshold of 0.02 au at the B3LYP/6−31G(d,p) level. Orbital 45 has the topology of a two-component torus link.

can be seen from the Poincaré-Hopf condition that creating one of each does not change the overall adherence to this rule. One can draw a (loose) analogy to the creation and subsequent annihilation of virtual pairs of nuclear particles. In our case, if the pair of BCP + RCP points is relatively close in space and in value of $\rho$(r), they will be prone to self-annihilate.[14] Looked at in this manner, the presence (or absence) of a BCP+RCP pair is merely the consequence of subtle changes in the electron density in the region of the pair and not a fundamental of the bonding of the molecule. Thus increasing the electron density in the C1−C9 region enables the creation of such a BCP+RCP pair, and it also increases the ELF V(C1,C9) threshold and the basin integration (0.98). Conversely, removing electron density might be expected to encourage any geometrically close BCP+RCP pair to self-annihilate.

The remaining entries in the Web-enhanced Table 1 provide support for this hypothesis. Thus the electron releasing **3**, X=$BH_2$ induces the creation of a BCP+RCP pair in the C1−C9 region, although the character of the ELF V(C1−B−C9) basin is tri- rather than disynaptic. The same effect is also obtained for the homotropylium analogue **1**, X=$BH_2$. The less electropositive **3**, X=$BF_2$ does not induce BCP+RCP creation (Figure 4 b,c), but it retains the trisynaptic nature of the valence basin. When the central atom in X is B or Al, the overall charge on the system is zero (it being zwitterionic), and the NICS(rcp) and $\Delta_r$ values suggest even more highly homoaromatic molecules. Indeed, even the anionic **3**, X=$BeH_2$ seems to sustain the effect.

The ELF(r)$_{C1-C9}$ basins also proved sensitive to substitution (Web-enhanced Table 1). If electron density is injected into this region, the vanishing threshold for the basin increases (toward that of a localized single bond). Conversely, electron withdrawal could inhibit even the formation of any basin in this region (e.g., X=$SiF_2$, $AlF_2$), and some basins have scarcely formed before they vanish at higher ELF(r) values (X=$BH_2$). These features suggest that analysis of ELF(r)$_{C1-C9}$ basins provides a more sensitive chemical probe than that provided purely by the presence or absence of bond critical points in the AIM analysis.

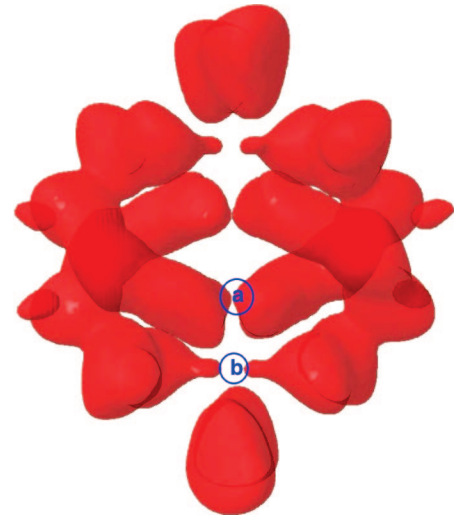



***Figure 6.*** ELF(r)$_\pi$=0.275 isosurface for **5**, X=$CH_2$, revealing that bifurcation of the endocyclic $\pi$-torus at point (a) is at the cusp, while that of the exocyclic $\pi$-torus (b) has already passed. The isosurface prior to bifurcation has the topology of a two-component torus link.

**2.3. Möbius Bishomoaromatics Based on 5.** The concept can also be extended to Möbius bis-homoaromaticity. The dication **5**, X=$CH_2$ comprises two 5$\pi$ monocationic components in which cyclic conjugation and hence aromaticity occurs across two $CH_2$ bridges. Having $D_2$-symmetry, it represents a double half-twist Möbius homoaromatic, for which a 4n+2 electron selection rule applies.[19] It sustains a large NICS(rcp) value of −15.8 ppm indicating a strong diatropic ring current, similar to that previously reported for a related nonhomoaromatic Möbius double half-twist system.[18,20] The neutral **5**, X=$BH_2$ is even more homoaromatic, with a shorter C−C homo bond length and a larger NICS(rcp) value. As before with **3**, the dicationic carbocyclic system (X=$CH_2$) shows no AIM bond-critical point, whereas the neutral boron system (X=$BH_2$) does. No ELF(r) all-electron disynaptic basin in the homobond region was located for X=$CH_2$, and so a dissection into its $\pi$-component was undertaken. The first five occupied MOs are in fact all $\pi$ in

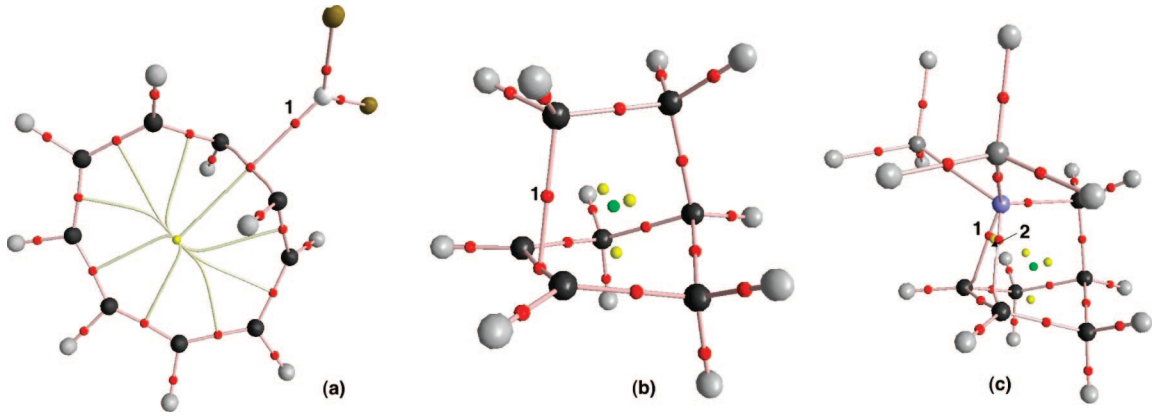
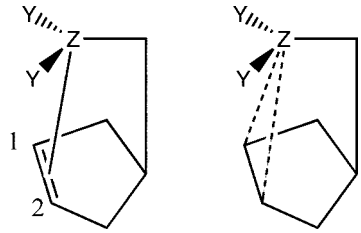



***Figure 7.*** Critical points for (a) **3**, X=$BeF_2$, (b) **6**, Z=C,Y=H. (c) **6**, Z=B,Y=$SiH_3$. The $\rho$(r) values (in e.$Å^{-3}$, B3LYP/6−31G(d,p)), at BCP point 1 (red) are 0.050, 0.085, and 0.105, respectively, and for RCP 2 (C), 0.105. Self-annihilation of points 1 and 2 in (c) results in point 1 in (b) remaining.

character (Figure 5). Noteworthy is the most stable of these (Figure 5e), which is cyclically continuous, and forms what is called a two component torus link that is identical in appearance to that described previously for the nonhomo forms.[19]

An ELF(r)$_\pi$ bufurcation takes place at an isosurface value of 0.275 (Figure 6) being that of the original $\pi$-torus at the two endocyclic regions of the homobond (the exocyclic torus having already bifurcated). The remaining $\pi$-torus does not fully bifurcate into basins until 0.986, a value even higher than that for benzene.[16] It is indeed tempting to conclude, by this measure at least, that a bis-homoaromatic system such as **5** is substantially more aromatic than even benzene!

As with **3**, a linking analysis for **5**, X=$CH_2$ reveals $L_k = 2\pi$, $T_w$=1.75, and $W_r = 0.25$, which again indicates relatively little of the double-half-twist of the ring has been transformed to writhe.




**6a** **6b**

**2.4. Comparisons with the Norbornyl Cation 6.** The specific example of **3**, X=$BeF_2$ gave the same number of BCPs as did X=$CH_2$, but they occurred in *different regions*. Whereas no BCP was present along the C1−C9 region for X=$CH_2$, two BCPs were located along C1-X and C9-X (Figure 1a). For X=$BeF_2$, one BCP is located in the C1−C9 region, and the second BCP lies along the path apparently connecting the first BCP to X (Figure 7a, point 1). This rather odd AIM-BCP behavior has in fact been computed[21] in another cationic system, the norbornyl cation **6** (also famously associated with Winstein). Two representations of the symmetrical species (in which the group $ZY_2$; Z=$C^+$, Y=H bridges symmetrically across the C1−C2 bond) have been argued over. Representation **6a** is that of a $\pi$-complex, in which the pair of $\pi$-electrons in the C1−C2 alkene donate into the vacant p-orbital on Z. The other, **6b**, is a nonclassical two-electron-three-center interpretation, in which (weaker) bonding is considered as occurring between each of C1 and C2 and Z. In terms of the AIM critical point analysis, **6a** results in a BCP along the C1−C2 path, and a second BCP occurs along the path connecting the first BCP and Z (Figure 7b). This is the result recovered for Z=$C^+$, Y=H. From our experiences above, one might expect a different result to be obtained if $ZY_2$ were to be made more electron releasing. Thus for the (neutral) system Z=B, Y=$SiH_3$, just such is computed. An additional BCP + RCP pair is created, and BCPs now occur along C1−C2, C1-Z, and C2-Z, and a RCP occurs at the centroid of this 3-membered ring, this effectively being representation **6b** (Figure 3c). Thus these two representations are therefore really just small, but subtle variations in the electron density topology, which differ only in whether a BCP+RCP pair is created or annihilated. Because this creation/annihilation only occurs when the critical points are close (it is estimated <0.2 Å),[14] this effect is only likely to be encountered for small, e.g. 3-membered rings; it no longer occurs for the larger 4-ring (**3**, X=$CH_2CH_2$).

## 3. Conclusions

The hitherto unexplored genre of Möbius homoaromaticity emerges as a new addition to the many diverse types of molecule that are regarded as having aromatic character. The effect appears equally prominent in cationic, neutral, and anionic molecules and also appears to sustain higher-order half-twists in the Möbius topologies, corresponding to e.g. bis-homoaromaticity. Unlike Hückel homoaromatics, the Möbius forms are intrinsically chiral (dissymmetric). Given the recent syntheses of a variety of Möbius single and higher order twist systems,[6] a speculation of whether any homo-Möbius aromatic might be synthetically accessible appears reasonable.

## 4. Computational Details

Calculations were performed at the B3LYP[22] DFT level and 6−31G(d,p) or aug-cc-pVTZ basis set level,[23] as implemented in the Gaussian 03 (revision E.01) program.[24] AIM critical points and the molecular graphs that map their connectivity were obtained by exporting a WFN file from Gaussian and importing into AIM2000.[25] The ring critical point coordinates so obtained were used to evaluate the NICS(rcp) values using Gaussian 03.[8]

ELF(r) and ELF(r)$_\pi$ cubes were calculated using Dgrid[26] via a FCHK file exported from Gaussian and visualized at differing isosurface values with VMD[27] or Jmol.[28] Basin integration was performed using TopMod.[29] The data files and coordinates are all available *via* the digital repository entries[30] to be found in the Web-enhanced Table 1.

## References


(1) Winstein, S. *J. Am. Chem. Soc.* **1959**, *81*, 6524–6525. DOI: 10.1021/ja01533a052.

(2) (a) Childs, R. F. *Acc. Chem. Res.* **1984**, *17*, 347–352. DOI: 10.1021/ar00106a001. (b) Cremer, D.; Kraka, E.; Slee, T. S.; Bader, R. F. W.; Lau, C. D. H.; Nguyen-Dang, T. T.; MacDougall, P. J. *J. Am. Chem. Soc.* **1983**, *105*, 5069–5075. DOI: ja00353a036. (c) Lepetit, C.; Silvi, B.; Chauvin, R. *J. Phys. Chem. A* **2003**, *107*, 464–473. DOI: 10.1021/jp0265211. (d) Cremer, D.; Reichel, F.; Krafka, E. *J. Am. Chem. Soc.* **1991**, *113*, 9459–9466. DOI: 10.1021/ja00025a006. (e) Geier, J. *J. Phys. Chem. A* **2006**, *110*, 9273–9281. DOI: 10.1021/jp061498f.

(3) Heilbronner, E. *Tetrahedron Lett.* **1964**, 1923–1928. DOI: 10.1016/S0040-4039(01)89474-0.

(4) (a) Zimmerman, H. E. *J. Am. Chem. Soc.* **1966**, *88*, 1564–1565. DOI: 10.1021/ja00959a052. (b) Zimmerman, H. E. *J. Am. Chem. Soc*. **1966**, *88*, 1566−1567. DOI: 10.1021/ja00959a053. (c) Zimmerman, H. E. *Tetrahedron* **1982**, *38*, 753–8. DOI: 10.1016/0040-4020(82)80155-5.

(5) (a) Rzepa, H. S. *Chem. Rev.* **2005**, *105*, 3697–3715. DOI: 10.1021/cr030092l. (b) Herges, R. *Chem. Rev.* **2006**, *106*, 4820–4842. DOI: 10.1021/cr0505425. (c) Ajami, D.; Oeckler, O.; Simon, A.; Herges, R. *Nature* **2003**, *426*, 819. (d) Ajami, D.; Hess, K.; Köhler, F.; Näther, C.; Oeckler, O.; Simon, A.; Yamamoto, C.; Okamoto, Y.; Herges, R. *Chem. Eur. J.* **2006**, *12*, 5434.

(6) (a) Kastrup, C. J.; Oldfield, S. V.; Rzepa, H. S. *Dalton Trans.* **2002**, 2421–2422. (b) Hall, D.; Rzepa, H. S. *Org. Biomol. Chem.* **2003**, *1*, 182–185. (c) Kui, S. C. F.; Huang, J.-S.; Sun, R. W.; Zhu, N.; Che, C. M. *Angew. Chem., Int. Ed.* **2006**, *45*, 4663–4666. (d) Park, J. K.; Yoon, Z. S.; Yoon, M. -C.; Kim, K. S.; Mori, S.; Shin, J.-Y.; Osuka, A.; Kim, D. *J. Am. Chem. Soc.* **2008**, *130*, 1824–1825. (e) Tunyogi, T.; Deák, A.; Tárkányi, G.; Király, P.; Pálinkás, G. *Inorg. Chem.* **2008**, *47*, 2049–2055. DOI:10.1021/ic702059v. (f) John, R. P.; Park, M.; Moon, D.; Lee, K.; Hong, S.; Zou, Y.; Hong, C. S.; Lah, M. S. *J. Am. Chem. Soc.* **2007**, *129*, 14142–14143. (g) Rzepa, H. S. *Org. Lett.* **2008**, *10*, 949–952. (h) Rzepa, H. S.; *Inorg. Chem*. **2008**,in press. (i) Stepien, M.; Latos-Grazynski, L.; Sprutta, N.; Chwalisz, P.; Szterenberg, L. *Angew. Chem., Int. Ed.* **2007**, *46*, 7869–7873. (j) Pacholska-Dudziak, E.; Skonieczny, J.; Pawlicki, Szterenberg, M. L.; Ciunik, Z.; Latos-Grazynski, L. *J. Am. Chem. Soc.* **2008**, *130*, 6182–6195.

(7) Mauksch, M.; Gogonea, V.; Jiao, H.; von, P.; Schleyer, R. *Angew. Chem. Int. Ed.* **1998**, *239*, 5–2397. DOI:10.1002/(SICI)1521-3773(19980918)37:17<2395::AID-ANIE2395>3.0.CO;2-W.

(8) (a) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema, H.; Nicolaas, J. R. *J. Am. Chem. Soc.* **1996**, *118*, 6317–6318. DOI: 10.1021/ja960582d. (b) Chen, Z.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R. *Chem. Rev.* **2005**, *105*, 3842–3888. DOI: 10.1021/cr030088+For a comparison of NICS with a range of other aromaticity indices, see: (c) Bultinck, P. *Faraday Discuss.* **2007**, *135*, 347–365. DOI: 10.1039/b609640a.

(9) Allan, C. S. M.; Rzepa, H. S. J. Org. Chem. **2008**, *73*, 6615–6622. DOI: 10.1021/j0801022b.

(10) (a) Winstein, S.; Kaesz, H. D.; Kreiter, C. G.; Friedrich, E. C. *J. Am. Chem. Soc.* **1965**, *87*, 3267–9. DOI: 10.1021/ja01092a060. (b) Aumann, R.; Winstein, S. *Tetrahedron Lett.* **1970**, *11*, 903–6. DOI: 10.1016/S0040-4039(01)97862-1.

(11) (a) Poater, J.; Duran, M.; Sola, M.; Silvi, B. *Chem. Rev.* **2005**, *105*, 3911–3947. DOI: 10.1021/cr030085x. (b) Bader, R. F. W. *Atoms in Molecules: a Quantum Theory*; Oxford University Press: Oxford, U.K., 1990. (c) Popelier, P. L. A. *Atoms in Molecules: an Introduction*; Prentice- Hall: London, 2000.

(12) (a) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–5403. DOI: 10.1063/1.458517. (b) Savin, A.; Jepsen, O.; Flad, J.; Andersen, O. K.; Preuss, H.; von Schnering, H. G. *Angew. Chem., Int. Ed.* **1992**, *31*, 187–188.

(13) (a) Lepetit, C.; Silvi, B.; Chauvin, R. *J. Phys. Chem. A* **2003**, *107*, 464–473. DOI: 10.1021/jp026521l. For a more general review of the electronic properties of homoaromaticity, see: (b) Cremer,D.; Childs, R. F.; Kraka, E. Chemistry of the Cyclopropyl Group;Rappaport, Z., Ed.; 1995; Vol. 2, pp 339−410.

(14) Castillo, N.; Matta, C. F.; Boyd, R. J. *Chem. Phys. Lett.* **2005**, *409*, 265–269. DOI: 10.1016/j.cplett.2005.04.088.

(15) (a) Savin, A.; Silvi, B.; Colonna, F. *Can. J. Chem.* **1996**, *74*, 1088–1096. DOI: 10.1139/v96-122. (b) Fuster, F.; Silvi, B. *Theor. Chem. Acc.* **2000**, *104*, 13–21. DOI: 10.1007/s002149900100. (c) Calatayud, M.; Andres, J.; Beltran, A.; Silvi, B. *Theor. Chem. Acc.* **2001**, *105*, 299–308. DOI: 10.1007/s002140000241. (d) Kohout, M.; Wagner, F. R.; Grin, Y. *Theor. Chem. Acc.* **2002**, *108*, 150–156. DOI: 10.1007/s00214-002-0370-x. (e) Cioslowski, J.; Matito, E.; Sola, M. *J. Phys. Chem. A* **2007**, *111*, 6521–6525. DOI: 10.1021/jp0716132.

(16) Santos, J. C.; Tiznado, W.; Contreras, R.; Fuentealba, P. *J. Chem. Phys.* **2004**, *120*, 1670–1673. DOI: 10.1063/1.1635799.

(17) Wannere, C. S.; Rzepa, H. S.; Rinderspacher, B. C.; Paul, A.; Schaefer, H. F.; Schleyer, P. v. R.; Allan, C. S. M., submitted for publication.

(18) (a) White, J. H. *Am. J. Math.* **1969**, *91*, 693–728. (b) Cālugāreanu, G. *Czech. Math. J.* **1961**, *11*, 588–625. (c) Fuller, F. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *68*, 815–819. (d) Pohl, W. *Indiana Univ. Math. J.* **1968**, *17*, 975–985.

(19) Rappaport, S.; Rzepa, H. S. *J. Am. Chem. Soc.* **2008**, *130*, 7613–7619. DOI: 10.1021/ja10438j.

(20) Rzepa, H. S. *Org. Lett.* **2005**, *7*, 4637–39. DOI: 10.1021/ol0518333.

(21) (a) Werstiuk, N. H.; Muchall, H. M. *THEOCHEM* **1999**, *463*, 225–229. DOI: 10.1016/S0166-1280(98)00625-3. (b) Werstiuk, N. H.; Muchall, H. M.; Noury, S. *J. Phys. Chem., A* **2000**, *104*, 11601–11605. DOI: 10.1021/jp001978l. (c) Firme, C. L.; Antunes, O. A. C.; Esteves, P. M. *J. Phys. Chem. A* **2008**, *112*, 3165–3171. DOI: 10.1021/jp710606n.

(22) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–5648. DOI: 10.1063/1.464304.

(23) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023. DOI: 10.1063/1.456153.

(24) 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.

(25) For computer programs, see: (a) Biegler-König, F. W.; Schönbohm, AIM2000. The program can be downloaded at http://www.aim2000.de/ (accessed 08/28/08). (b) Keith, T. A., AIMAll (Version 08.04.21), 2008. The program can be downloaded at http://aim.tkgristmill.com (accessed 08/28/08).

(26) For Computer Program, see: (a) Kohout, M. DGrid, version 4.3; 2008.

(27) For Computer Program, see: (a) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38, Version 1.8.6, 2007. The program can be downloaded at http://www.ks.uiuc.edu/Research/vmd/ (accessed 08/28/08).

(28) For Computer Program, see: Hanson, R., Ed.; 2008. The program can be downloaded at http://jmol.sf.net/ (accessed 08/28/08).

(29) For Computer Program, see: Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. TopMoD package; Universite Pierre et Marie Curie, 1997. The program can be downloaded athttp://www.lct.jussieu.fr/pagesperso/silvi/topmod.html (accessed 08/28/08).

(30) Rzepa, H. S.; Downing, J.; Tonge, A.; Cotterill, F.; Harvey, M. J.; Murray-Rust, Morgan, P.; Day, N. J. Chem. Inf. Model. **2008**, *48*, 1571–1581. DOI: 10.1021/ci7004737.

CT8001915

JCTC Journal of Chemical Theory and Computation

# Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions

Yan Zhao and Donald G. Truhlar*

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

Received June 25, 2008

**Abstract:** The hybrid meta density functionals M05-2X and M06-2X have been shown to provide broad accuracy for main group chemistry. In the present article we make the functional form more flexible and improve the self-interaction term in the correlation functional to improve its self-consistent-field convergence. We also explore the constraint of enforcing the exact forms of the exchange and correlation functionals through second order (SO) in the reduced density gradient. This yields two new functionals called M08-HX and M08-SO, with different exact constraints. The new functionals are optimized against 267 diverse main-group energetic data consisting of atomization energies, ionization potentials, electron affinities, proton affinities, dissociation energies, isomerization energies, barrier heights, noncovalent complexation energies, and atomic energies. Then the M08-HX, M08-SO, M05-2X, and M06-2X functionals and the popular B3LYP functional are tested against 250 data that were not part of the original training data for any of the functionals, in particular 164 main-group energetic data in 7 databases, 39 bond lengths, 38 vibrational frequencies, and 9 multiplicity-changing electronic transition energies. These tests include a variety of new challenges for complex systems, including large-molecule atomization energies, organic isomerization energies, interaction energies in uracil trimers, and bond distances in crowded molecules (in particular, cyclophanes). The M08-HX functional performs slightly better than M08-SO and M06-2X on average, significantly better than M05-2X, and much better than B3LYP for a combination of main-group thermochemistry, kinetics, noncovalent interactions, and electronic spectroscopy. More important than the slight improvement in accuracy afforded by M08-HX is the conformation that the optimization procedure works well for data outside the training set. Problems for which the accuracy is especially improved by the new M08-HX functional include large-molecule atomization energies, noncovalent interaction energies, conformational energies in aromatic peptides, barrier heights, multiplicity-changing excitation energies, and bond lengths in crowded molecules.

## 1. Introduction

The development of new and better exchange-correlation functionals for density functional theory (DFT) is "promising and charming".[1] We classify functionals as local and nonlocal. In the classification used here, at a given point in space, local density functionals depend on at most the spin densities and their derivatives and spin kinetic energy density at that point in space; nonlocal functionals involve an integral over all space. The only widely studied (to date) method to include nonlocality is to incorporate Hartree−Fock (HF) exchange; functionals involving HF exchange are called hybrid. In order of increasing complexity and accuracy, the three types of functionals that we classify as local are the local spin density approximation (LSDA),[2] generalized gradient approximation (GGA),[3−8] and meta-GGAs.[9−13] Nonlocal functionals include hybrid GGAs and hybrid meta

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

10.1021/ct800246v CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

functionals. Hybrid GGAs (which include nonlocal HF exchange) have better performance for general-purpose applications in chemistry than the local functionals. One hybrid GGA, namely B3LYP,[4,5,14,15] has become extraordinarily popular.[16] Some later hybrid GGAs, though, such as mPW1PW,[17] PBEh,[18] and B97−3[19] (and, for barriers and noncovalent interactions, MPW1K[20]), have better performance than B3LYP. Hybrid meta density functionals[9,21−28] in which the energy depends on the occupied orbitals not only through the HF exchange terms (as in hybrid GGAs) but also through the noninteracting spin kinetic energy densities[29−32] (as in meta-GGAs) have been shown to be capable of even better performance than hybrid GGAs.[13,23−28] To distinguish them from some new developments mentioned below, conventional hybrid GGAs and hybrid meta functionals may be called global hybrid functionals and global hybrid meta functionals, respectively.

Recently, evidence disparaging the performance of popular density functionals for many areas in chemistry has been presented by many research groups.[33−49] In order to improve the performance of conventional density functionals, besides the design and semiempirical fitting conventional exchange-correlation functionals, four newer approaches have been proposed, namely.

1) DFT-D: DFT-D augments the DFT energy by a damped dispersion term (in the functional or added to the energy as in combined quantum mechanical and molecular mechanical methods) that yields the correct asymptotic form $-C_6R^{-6}$ (plus possibly higher order terms, if the multipole expansion is not truncated at the first term) of the interatomic or intermolecular interaction potential.[50−65] Some recent successful DFT-D functionals are TPSS-D,[53] B97-D,[52,64] and DF07.[65]

2) Range-separated hybrid (RSH) functionals: The RSH approach was first proposed by Savin;[66] in this approach the Coulomb operator is partitioned into long-range and short-range parts, and different treatments are employed for the long-range and short-range operators. Some recent functionals of this type are HSE03,[67] CAM-B3LYP,[68] RSHXPBE,[69] LC-$\omega$PBE,[70] LCgau-BOP,[71] PBE/CCSD,[72] and $\omega$B97X.[73]

3) Local hybrid functionals: the amount of exact HF and DFT exchange in the local hybrid functional, unlike the global hybrid functional, varies according to the local properties of each system.[74] Some developments in refining the local mixing functions have been reported recently.[75−80]

4) Doubly hybrids: A doubly hybrid functional is a hybrid of a global hybrid functional with correlation contributions from unoccupied orbitals. We developed several doubly hybrid models by empirically mixing correlated wave function methods and density functional methods.[81,82] In our published models, we used the HF orbitals for the unoccupied orbitals, although in unpublished work carried out at the time we found that we got similar results for the systems studied by using Kohn−Sham orbitals. In the recent B2PLYP[83] and mPW2PLYP[84] doubly hybrid functionals, Grimme et al. employed the Kohn−Sham unoccupied orbitals to calculate the second-order Møller−Plesset-type perturbation theory correction. (The first-order contribution, which is nonzero when one uses Kohn−Sham orbitals, was omitted.) More recently, Tarnopolsky et al.[85] reoptimized the parameters in B2PLYP and mPW2PLYP for thermochemical kinetics, resulting in the B2K-PLYP and mPW2K-PLYP functionals. Benighaus et al.[86] also optimized a doubly hybrid functional, called B2-P3LYP, and they noted that one major drawback of B2-PLYP or B2-P3LYP is their fifth-order scaling with respect to system size. They proposed a fourth-order scaling doubly hybrid functional, namely B2-OS3LYP, by just retaining the opposite-spin component of the Møller−Plesset-type of correlation energy. Note that these approaches go beyond the hybrid and meta approaches in that they introduce a dependence not only on occupied orbitals but also on the unoccupied-orbital space; as such they are sometimes called fifth-rung[87,88] DFT.

Note that some functionals are developed by combining two of the four approaches, such as the RSE-MP2 functional,[89] which is developed by combining approaches 2) and 4). The B2-PLYP-D and mPW2-PLYP-D functionals were developed by combining approaches 1) and 4).[90]

Among these four kinds of treatments, the range-separated hybrids have a computer cost for molecules that is approximately the same as that of global hybrids, and the HSE functional of Heyd et al.[67] has much improved computational cost for solids. However, the goal of the present study is different from the above-mentioned lines of research. The question we want to address in the present study is the following: *without these nonconventional treatments, how accurate can a global hybrid meta functional be for main-group chemistry*. In particular, what is the limit of accuracy of the global hybrid meta functional form for a combination of three important areas of chemistry:

TC main-group thermochemistry

BH barrier heights

NC noncovalent interactions

We may compare this effort to Becke's work in 1997 and 1998 in which[91,92] he varied parameters to determine the limit of accuracy attainable for TC-type predictions by the hybrid GGA functional form. The final resulting functional, B98, remains to this day a good representative of the best that one can do for main-group thermochemistry with a hybrid GGA.

Recently we have shown that our global hybrid meta M05-2X[27] functional performs well for many problematic systems,[93−100] and this good performance has been confirmed by recent studies from other groups for organic energies,[101] for conjugated addition reaction energetics,[102] for peptides containing an aromatic ring,[103] for the conformational energetics of isocohichine,[104] for excited states of stacked nucleobases,[105] for near-edge X-ray and optical absorption spectra of liquid water,[106] and for the contribution of dipole−dipole interactions to the stability of the collagen triple helix.[107] It is encouraging that these assessments are diverse and outside the training set of M05-2X, showing the transferability of the optimized parameters in the M05-2X functional. Subsequent work showed even an better than average performance by the M06-2X functional,[13,28] which can be considered to be an improved version of M05-2X. M06-2X has been successfully employed to explain the unusual temperature dependence of an atmospherically

***Table 1.*** Basis Set and Geometries

| databases | ref | basis sets | geometries |
|---|---|---|---|
| | | Training Sets | |
| MGAE109 | 27 | MG3S | QCISD/MG3 |
| IP13 | 26,27,113,114 | MG3S | QCISD/MG3 |
| EA13 | 26,27,113,114 | MG3S | QCISD/MG3 |
| PA8 | 93 | 6−311+G(2df,2p) | MP2(full)/6−31G(2df,p) |
| ABDE4 | 11,27,37 | 6−311+G(3df,2p) | B3LYP/6−31G(d) |
| $\pi$IE3 | 11,27 | 6−311+G(2df,2p) | MP2/6−31+G(d,p) |
| PA-CP5 | 93 | 6−311+G(2df,2p) | MP2/6−31+G(d,p) |
| PA-SB5 | 93 | 6−311+G(2df,2p) | MP2/6−31+G(d,p) |
| DBBH76 | 27,82,115 | MG3S | QCISD/MG3 |
| NCCE31 | 113,116 | MG3S | MC-QCISD/3 |
| AE17 | 28,117 | MQZVP, MG3S, aug-cc-pVQZ[a] | N. A. |
| | | Test Sets | |
| G3-3AE75 | 22,33, this work | MG3SXP, 6−311++G(3df,3pd) | B3LYP/6−31G(2df,p) |
| LMAE14 | 118, this work | 6−311+G(3df,2p) | M06-L/6−311+G(2df,2p) |
| IE34 | 119,45 | MG3S | B3LYP/TZV(d,p) |
| S22 | 120 | 6−311+G(3df,2p) | WFT |
| APCE5 | 121 | 6−311+G(3df,2p) | RI-MP2/cc-pVTZ |
| UUU7 | 122 | 6−311+G(3df,2p) | RI-MP2/cc-pVTZ |
| BBH7/08 | compiled in this work | 6−311+G(3df,2p) | Opt[b] |
| MGBL24 | 11,28, this work | MG3S | Opt[b] |
| CID15 | 123,124 | MG3S | Opt[b] |
| F38/06 | 28 | MG3S | Opt[b] |
| MGMCEE9 | compiled in this work | aug-cc-pVQZ | experiment + opt[c] |

[a] MQZVP is used for training, and MG3S and aug-cc-pVQZ are used for testing. [b] Opt denotes that the geometry is reoptimized for each density functional tested. [c] The molecules for vertical transitions are calculated at fixed geometries as specified in Section 2.2.12, whereas the molecules for adiabatic transitions are reoptimized for each density functional tested with the 6−311+G(2df,2p) basis set.

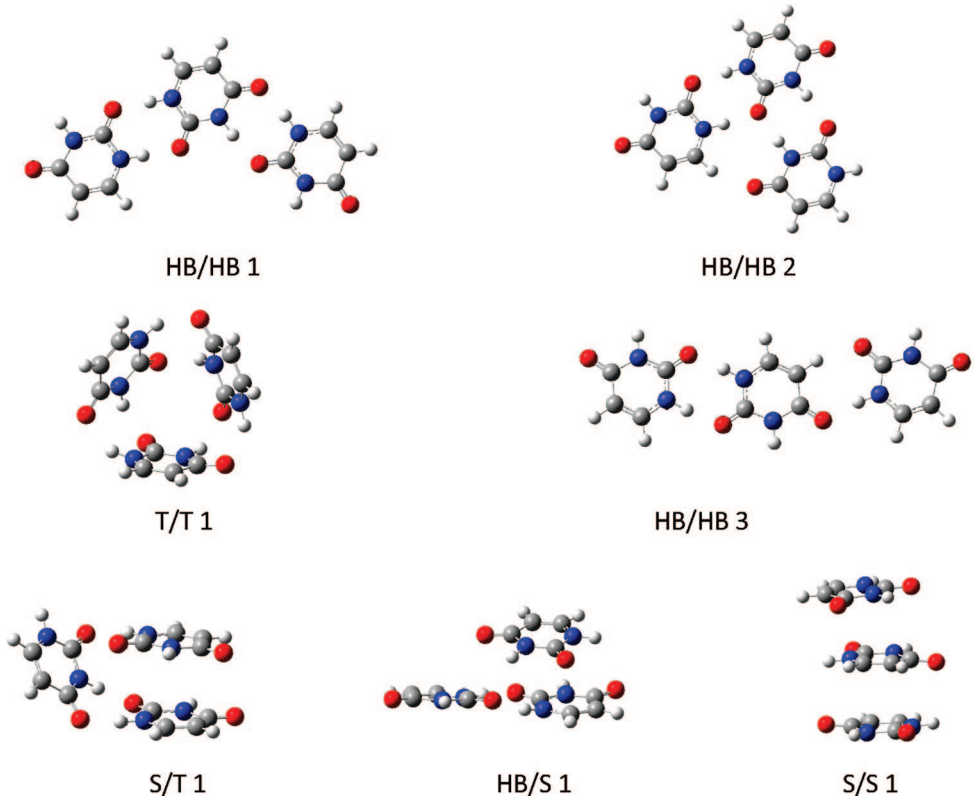



***Figure 1.*** Structures of uracil trimers in the UUU7 database.

important reaction,[108] to study steric isotope effects,[109] to investigate the structures and potential energy surface of coronene dimers,[110] and to calculate host−guest interactions in supramolecular complexes in a hydrocarbon nanoring[111] and concave-convex $\pi...\pi$ interaction in buckyball tweezers.[112] Note that these applications are beyond the reach of the popular functionals. Encouraged by these successes, we investigate in the present study the extent to which further improvements can be achieved using a more flexible functional form (although still of the hybrid meta type), and we call the new functional M08-HX, where "X" is our usual abbreviation for Hartree−Fock exchange, and M08-HX

denotes "Minnesota 2008 high-X". Using the same functional forms as M08-HX, we optimize another functional in which we enforced the gradient expansion to the second order, and we call this functional M08-SO, where SO denotes "second order".

The paper is organized as follows. Section 2 presents our databases. Section 3 gives computational details. Section 4 discusses the theory and parametrization of the new functionals. Section 5 presents results and discussion not only for the TC, BH, and NC areas mentioned above but also for bond lengths, vibrational frequencies, and multiplicity-changing excitation energies. Section 6 concludes the paper.

## 2. Databases

All databases used in this article are listed in Table 1. Table 1 also presents the references for the databases[11,22,26−28,33,37,45,82,93,113−124] and the basis sets and geometries employed here for each database. The entries in Table 1 are explained in this section. One important difference between the energetic data in the present article and the energetic data used in many other studies is that we exclusively use clamped-nuclei energies (such as equilibrium dissociation energies, $D_e$, adiabatic clamped-nuclei IPs, or classical barrier heights) rather than 0 K data (such as ground-state dissociation energies, $D_0$, which include zero point vibrational energy) or 298 K data (such as finite-temperature enthalpies of activation or heats of formation that also include thermal vibrational−rotational energies). Using 0 or 298 K data provides a combined test of the ability to predict electronic, vibrational, and rotational energies (rotational energies depend on geometries). In contrast, by making databases of our best estimates of clamped-nuclei energies (which consist of the electronic energy including nuclear repulsion), we obtain pure tests of Born−Oppenheimer electronic energies. We test vibrational frequencies and bond lengths with separate databases.

**2.1. Training Sets.** We used the same training sets for M08-HX and M08-SO as for M06-2X,[28] including the MGAE109 database of 109 main-group atomization energies (equilibrium dissociation energies for complete dissociation to ground-state atoms),[27] the IP13 database of 13 ionization potentials, the EA13 database of 13 electron affinities, the PA8 database of 8 proton affinities, the ABDE4 database of four alkyl bond dissociation energies, the DBBH76 database of 76 diverse barrier heights, the $\pi$IE3 database of three isomeric energy differences between allene and propyne as well as higher homologues, the PA-CP5/06 database of the proton affinities of five conjugated polyenes, the PA-SB5/06 database of the proton affinities of the five conjugated Schiff bases, the NCCE31 database of 31 noncovalent interaction energies (6 hydrogen bonds, 7 charge transfer complexes, 6 dipole interactions, 7 weak interactions, and 5 $\pi$...$\pi$ interactions), and the AE17 database of 17 nonrelativistic atomic energies[117] for the atoms from H to Cl.

**2.2. Test Sets.** The training sets for M08-HX and M08-SO consist of mainly small molecules and complexes. We test the functionals outside the training sets and compare the results in all cases with the popular functional B3LYP and the earlier M05-2X functional and in some cases with

***Table 2.*** Optimized Parameters for M08-H and M08-SO

| | M08-HX | | | | M08-SO | | | |
|---|---|---|---|---|---|---|---|---|
| i | $a_i$ | $b_i$ | $c_i$ | $d_i$ | $a_i$ | $b_i$ | $c_i$ | $d_i$ |
| 0 | 2.7925837E+00 | −1.7925858E+00 | 1[a] | 1.3812334E+00 | −8.0741559E-01 | 1.8074156E+00 | 1[a] | 1[a] |
| 1 | −1.9834852E+01 | 1.9428586E+01 | −4.0661387E-01 | −2.4683806E+00 | −1.3459249E+01 | 1.2621657E+01 | 0[a] | −4.4117403E+00 |
| 2 | −2.6254749E+01 | 2.5666211E+01 | −3.3232530E+00 | −1.1901501E+01 | 8.6903055E+01 | −8.7603211E+01 | −3.9980886E+00 | −6.4128622E+00 |
| 3 | 1.9127062E+01 | −1.1553206E+01 | 1.5540980E+00 | −5.4112667E+01 | 1.4748300E+02 | −1.4416820E+02 | 1.2982340E+01 | 4.7583635E+01 |
| 4 | 7.2675747E+01 | −7.4387668E+01 | 4.4248033E+01 | 1.0055846E+01 | −1.2437471E+02 | 1.0810751E+02 | 1.0117507E+02 | 1.8630053E+02 |
| 5 | 1.2315639E+02 | −1.7176051E+02 | −8.4351930E+01 | 1.4800687E+02 | −2.2817757E+02 | 2.0208604E+02 | −8.9541984E+01 | −1.2800784E+02 |
| 6 | 1.4940250E+02 | −1.4357663E+02 | −1.1955581E+02 | 1.1561420E+02 | 3.7681592E+01 | 3.7152154E+01 | −3.5640242E+02 | −5.5385258E+02 |
| 7 | 4.8802514E+01 | 7.5540493E+01 | 3.9147081E+02 | 2.5591815E+02 | 4.0531053E+01 | 4.6579309E+01 | 2.0698803E+02 | 1.3873727E+02 |
| 8 | 1.0114021E+01 | −1.9623400E+01 | 1.8363851E+02 | 2.1320772E+02 | −1.5650903E+01 | −9.3366569E+01 | 4.6037780E+02 | 4.1646537E+02 |
| 9 | −1.3616112E+01 | −1.2504017E+02 | −6.3268223E+02 | −4.8412067E+02 | 2.5703906E+01 | −1.3556484E+02 | −2.4510559E+02 | −2.6626577E+02 |
| 10 | −2.9429067E+01 | 3.4724447E+01 | −1.1297403E+02 | −4.3430813E+02 | 3.6249816E+00 | 4.8345920E+01 | −1.9638425E+02 | 5.6676300E+01 |
| 11 | 2.6963722E+01 | 2.9292867E+01 | 3.3629312E+02 | 5.6627964E+01 | 2.0273888E+01 | 2.5334189E+01 | 1.1881459E+02 | 3.1673746E+02 |
| $X$ | 52.23 | 56.79 | | | | | | |

[a] Determined by constraints, as explained in Section 4.4.

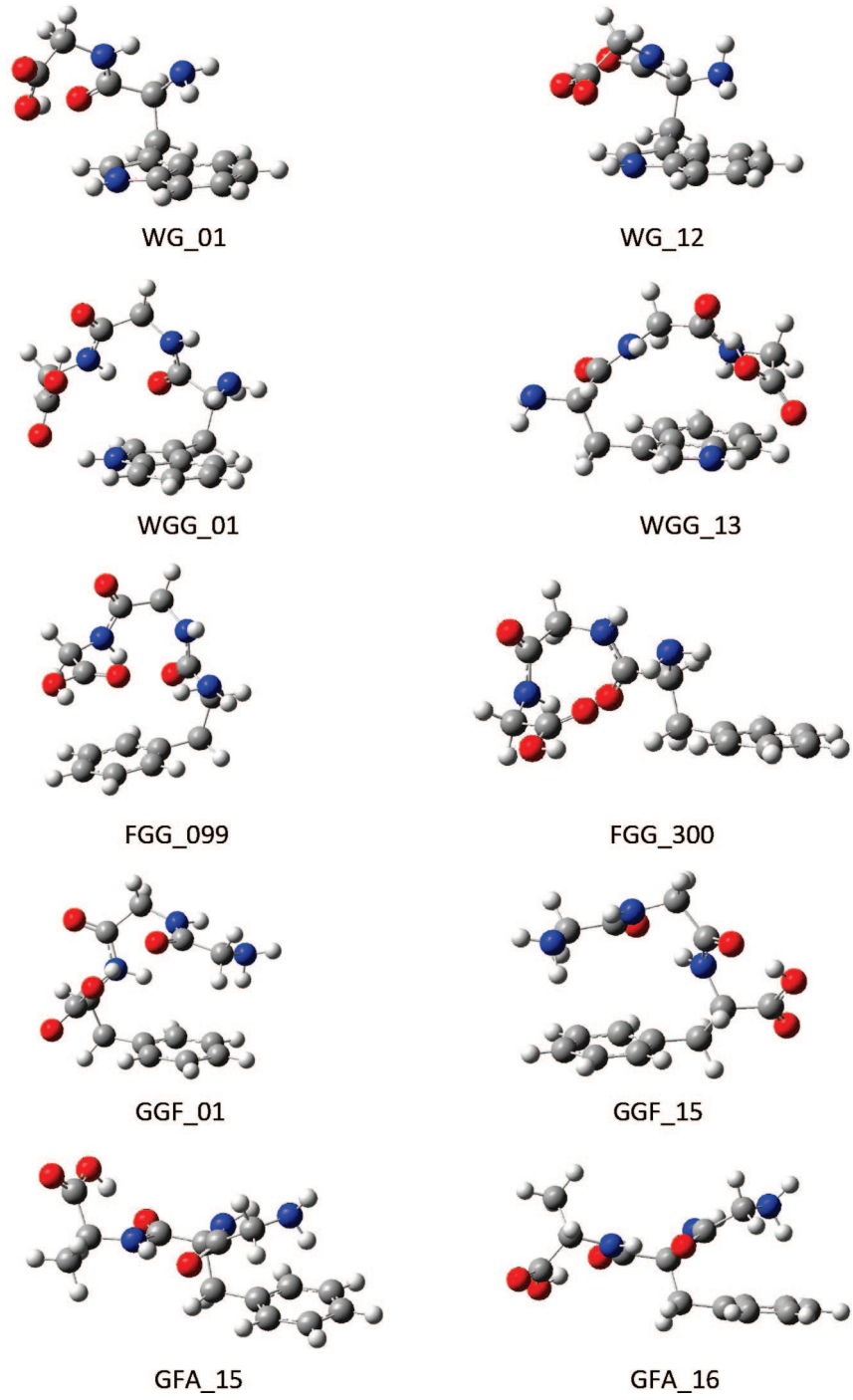



***Figure 2.*** Structures of small peptides in the APCE5 database.

additional selected functionals as well. As indicated in each case, some of the comparison results are taken from the literature, and many others are newly computed especially for this article. The additional databases used for testing are explained next.

*2.2.1. G3-3AE75.* G3-3AE75 is a database of 75 atomization energies for the molecules in the G3-3 set of Curtiss et al.[33] We used experimental standard enthalpies of formation (at 298 K)[33] and scaled B3LYP/6−31G(2df,p) harmonic zero-point vibration energies (with a scaling factor of 0.9854)[22] and thermal contributions to obtain reference clamped-nuclei experimental atomization energies. This database is given in the Supporting Information.

*2.2.2. LMAE14.* LMAE14 is a database of 14 large-molecule atomization energies for molecules that have 56 or more valence electrons. Most of the molecules in this set are not feasible for G3 methods.[118] We used experimental standard enthalpies of formation (at 298 K)[118] and scaled M06-L/6−311+G(2df,2p) zero-point vibration energies (with a scaling factor of 0.98) and thermal contributions to obtain reference experimental atomization energies.

*2.2.3. IE34.* IE34 is a benchmark database of 34 organic isomerization energies compiled by Jorgensen et al.[119,125] Grimme et al.[45] found that four experimental data in this database are not reliable as compared to high-level CCSD(T)

calculations. We use the reference data of Grimme et al.[45] for this database.

*2.2.4. S22 Database.* The S22 database is a data set of 22 weakly bonded molecular complexes of biological importance. This database was developed by Jurecka et al.,[120] who divided the S22 set into three subsets, namely, 7 hydrogen bonded complexes, 8 dispersion-dominated complexes, and 7 mixed complexes. The reference interaction energies for the S22 data set were calculated[120] by the following scheme

$$\Delta E^{CCSD(T)}CBS = \Delta E^{MP2}CBS + (\Delta E^{CCSD(T)} - \Delta E^{MP2})_{small\ basis} \quad (1)$$

where a complete basis set (CBS) limit CCSD(T) interaction energy is approximated by a CBS MP2 interaction energy plus a difference between CCSD(T) and MP2 interaction energies ($\Delta E^{CCSD(T)} - \Delta E^{MP2}$) evaluated with a relatively small basis set that was specifically designed[126,127] for this purpose. The best estimates of the interaction energies in the S22 database were taken from the paper by Jurecka et al.[120]

*2.2.5. UUU7.* UUU7 is a benchmark database of 7 noncovalent interaction energies in uracil trimers. The structures for the 7 trimers in UUU7 are shown in Figure 1. The reference data are based on the estimated CCSD(T)/CBS (eq 1) results of Kabelác et al.[122] We used the same name convention as in ref 122 to label these trimers.

*2.2.6. APCE5.* APCE5 is a benchmark database of 5 aromatic peptide conformational energies in 5 small peptides containing an aromatic side chain, taken from a recent benchmark database compiled by Valdes et al.[121] In particular, APCE5 includes the energy gaps between the highest-energy conformer and the lowest-energy conformer at the estimated CCSD(T)/CBS level for the WG, WGG, FGG, GGF, and GFA peptides containing phenylalanine (F), glycine (G), trypophan (W), and alanine (A). The structures of the 5 pair of peptides are shown in Figure 2, and we use the same name convention in ref 121. The reference conformational energy gaps for the five small peptides are calculated from the estimated CCSD(T)/CBS (eq 1) results of Valdes.[121]

*2.2.7. BBH7/08.* BBH7/08 is a new (2008) database of 7 diverse benchmark barrier heights in 5 reactions, in particular

$$CH_3 + CH_4 \rightarrow CH_4 + CH_3 \quad (R1)$$

$$HCC + HCCH \rightarrow HCCH + CCH \quad (R2)$$

$$OH^- + CH_3OH \rightarrow CH_3OH + OH^- \quad (R3)$$

$$HCC + H_2 \rightarrow HCCH + H \quad (R4)$$

$$O + CH_4 \rightarrow OH + CH_3 \quad (R5)$$

The reference data for reactions R1, R2, and R4 are based on W1[128] calculations, and they were taken from our previous study.[129] The reference classical barrier height for R3 was taken from a focal point calculation of Gonzales et al.[130] The reference forward and reverse classical barrier heights for R5 are based on the CCSD(T)/CBS calculations by Troya.[131]

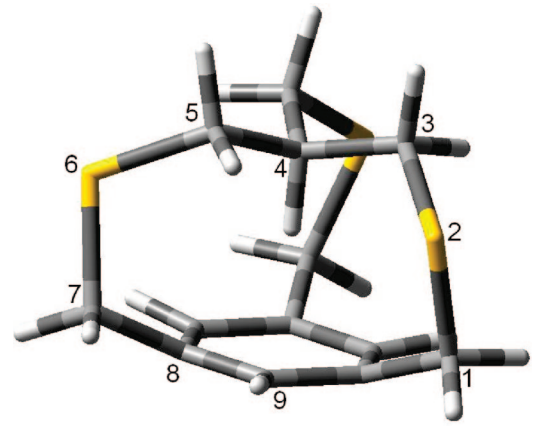



***Figure 3.*** Structure of 2,6,15-trithia-*in*-[$3^{4,10}$][7]metacyclophane.

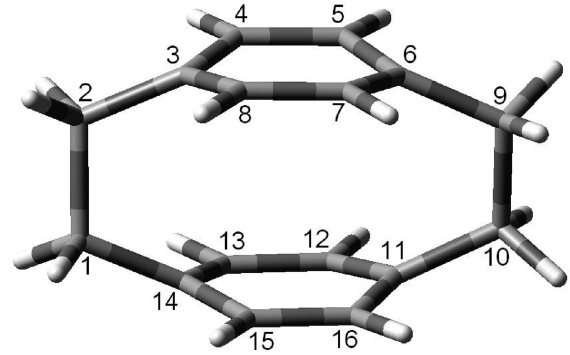



***Figure 4.*** Structure of [2.2]paracyclophane.

*2.2.8. MGBL24.* MGBL24 is a database of 24 bond lengths in 20 molecules. This database is based on our previous MGBL19 database.[11] We augmented the MGBL19 database by including three main-group metal dimers ($Li_2$, $Na_2$, $Al_2$), an open-shell molecule (BN), and a high-coordination molecule ($SF_6$). The reference data were taken from Computational Chemistry Comparison and Benchmark Database[132] and from Handy and Tozer.[133]

*2.2.9. CID15.* CID15 is a database for 15 internuclear distances in two cyclophanes, namely 2,6,15-*in*-trithia[$3^{4,10}$]-[7]metacyclophane and [2,2]paracyclophane. The structures of these molecules are shown in Figures 3 and 4. The reference data were taken from Pascal et al.[123] and Grimme.[124]

*2.2.10. F38/06.* F38 is a database of 38 harmonic frequencies compiled in a previous study,[28] which consists of the F36/06 database[11] plus the harmonic frequencies[134] of the OH and $Cl_2$ molecules.

*2.2.11. MGMCEE9.* MGMCEE11 is a database of 9 main-group electronic excitation energies for transitions to the lowest-energy excited states with a multiplicity different from the ground state, including two atoms (Be and Mg)[28] and seven molecules. Among the seven molecules, we have vertical excitations for five molecules (BeH, CO, $H_2CO$, $H_2O$, and $N_2$) at fixed geometries and adiabatic excitations for two molecules ($NO_2$ and SiO). The best estimate of the vertical excitation and the geometry for BeH ($r_e = 1.326903$) is from the FCI calculation of Pitrach-Ruiz et al.[135] The geometries (in Å and deg) for CO ($r_{CO} = 1.128$), $H_2CO$ ($r_{CO} = 1.203$,

$r_{CH} = 1.102$, $\theta_{HCO} = 121.9$), $N_2$ ($r_{NN} = 1.098$), and $H_2O$ ($r_{OH} = 0.957$, $\theta_{HOH} = 104.5$) are taken from Handy and Tozer.[133] The reference vertical excitation energies for $N_2$, CO, and $H_2CO$ are from experiments,[136−138] whereas the reference vertical excitation energy for $H_2O$ is determined in the present study by using high-level WFT calculations. The reference adiabatic excitation energy for SiO is taken from NIST Chemistry Webbook,[134] whereas that for $NO_2$ is taken from a benchmark calculation of Bera et al.[139]

## 3. Computational Methods

**3.1. Geometries and Basis Sets.** The basis sets and geometries used for the training sets and test sets in the present article are listed in Table 1. The 6−311+G(2df,2p),[140−142] 6−311+G(3df,2p),[140−142] 6−311+G(3df,3pd),[140−142] DIDZ (short name of 6−31+G(d,p)),[142−144] MG3S,[145] MQZVP,[28,146] and aug-cc-pVQZ[147] basis sets are explained elsewhere. A new basis set, MG3SXP, is used for some of the calculations in this study. The MG3SXP (where XP denotes "extra polarization") basis differs from the MG3S[114] basis set in the same way that G3LargeXP differs from G3Large,[148] in particular, the 2df polarization functions of MG3S on Li−Ne are replaced by a 3df set, and the 3d2f polarization functions on Al−Ar are replaced by 4d2f, where the polarization functions are those recommended by Curtiss et al.[148]

**3.2. Counterpoise Corrections.** For the noncovalent complexes in S22, we performed calculations with and without the counterpoise (CP) corrections[149,150] for basis set superposition error (BSSE). The results for the NCCE31 database are CP corrected, whereas the results for the UUU7 database are CP uncorrected.

**3.3. Spin−Orbit Energy.** Except for the AE17 database (for which the reference data are from high-level nonrelativistic WFT calculations) and except when explicitly indicated otherwise in Section 5.2.1, the spin−orbit energy is added for all species for which it is nonzero. A complete list of spin−orbit energies used for calculations in this article can be found elsewhere.[151]

**3.4. Software.** All DFT calculations in this article were performed with a locally modified version of the *Gaussian03* program.[152,153] The high-level WFT calculations that are used to determine the reference vertical excitation energy of $H_2O$ are performed with the *NWChem* program.[154]

**3.5. Excitation Energies.** The multiplicity-changing excitation energies are not calculated with time-dependent DFT but by taking the energy difference between ground states of different multiplicity.

## 4. Theory and Parametrization

The local parts of the M08-HX and M08-SO functionals depend on six variables: up-spin and down-spin densities ($\rho_\alpha$ and $\rho_\beta$), their density gradients ($\nabla\rho_\alpha$ and $\nabla\rho_\beta$), and spin kinetic energy densities ($\tau_\alpha$ and $\tau_\beta$).

**4.1. M08-Type Exchange Functional.** The spin-scaling relation[155] for exchange energy allows us to explain the exchange functionals by considering the exchange functional for a spin-unpolarized system for which $\rho = 2\rho_\alpha = 2\rho_\beta$. In the meta-GGA framework, the exchange energy of a spin-unpolarized system can be written as

$$E_x^{GGA}[\rho] = \int d^3r \rho \varepsilon_x^{LDA}(\rho) F_x(s,\tau) \quad (2)$$

where $s$ is the dimensionless reduced density gradient given by

$$s = |\nabla\rho| / [2(3\pi^2)^{1/3}\rho^{4/3}] \quad (3)$$

where $\varepsilon_x^{LDA}$ is the local density approximation[2] for the exchange energy per particle, and $F_x(s,\tau)$ is the meta-GGA exchange enhancement factor.

A key element in the exchange functional is the Taylor series coefficient defined by

$$\mu \equiv \lim_{s\to 0}\left(\frac{1}{2}\frac{d^2F_x}{ds^2}\right) \quad (4)$$

The accurate value of this coefficient is well-known, and we will call it $\mu^{GE}$ where GE denotes gradient expansion; the accurate value is $10/81 = 0.1235$.[156]

The M08-type exchange functional form is based on our SOGGA exchange functional,[8] which is a half-and-half-mix of the PBE[7] and RPBE[157] exchange functional forms, that is

$$E_x^{SOGGA} = \int dr \rho \varepsilon_x^{LDA}(0.5F_x^{PBE} + 0.5F_x^{RPBE}) \quad (5)$$

where $F_x^{PBE}$ is the enhancement factor for the PBE[7] exchange, and $F_x^{RPBE}$ is the enhancement factor for the RPBE[157] exchange:

$$F_x^{RPBE} = 1 + \kappa_2(1 - e^{-\mu_2 s^2/\kappa_2}) \quad (6)$$

In the present work eq 5 is first generalized to the following local form:

$$E_x^{M08\text{-}Loc} = \int d^3r \rho \varepsilon_x^{LDA}\{f_1(w)F_x^{PBE} + f_2(w)F_x^{RPBE}\} \quad (7)$$

$f_1(w)$ and $f_2(w)$ are the kinetic-energy-density enhancement factors

$$f_1(w) = \sum_{i=0}^{11} a_i w^i \quad (8)$$

$$f_2(w) = \sum_{i=0}^{11} b_i w^i \quad (9)$$

where the variable $w$ is a function of $y$, and $y$ is a function of the kinetic energy density $\tau$ (which equals $2\tau_\alpha$ or $2\tau_\beta$ for a spin-unpolarized system) and density $\rho$

$$w = (y-1)/(y+1) \quad (10)$$

$$y = \tau^{UEG}/\tau \quad (11)$$

where $\tau^{UEG}$ is the Thomas-Fermi[158,159] kinetic energy density for a uniform electron gas (UEG)

$$\tau^{UEG} = \frac{3}{10}(3\pi^2)^{2/3}\rho^{5/3} \quad (12)$$

and

$$\tau = \frac{1}{2}\sum_{i=1}^{n}|\nabla\varphi_i|^2 \quad (13)$$

where $\varphi_i$ is a generalized Kohn−Sham orbital (also called a Hartree−Fock Kohn−Sham orbital), and $n$ ($\equiv n_\alpha + n_\beta$) is the number of occupied orbitals. (Note that some authors (including ref 160) define $\tau$ without the factor of 1/2.)

For a slowly varying density, $\tau$ has the gradient expansion (GE)[160]

$$\tau = \tau^{\mathrm{UEG}} + \frac{1}{72}\frac{|\nabla\rho|^2}{\rho} + O(\nabla^2) \quad (14)$$

Using eqs 10−14, we can derive a GE approximation for $y$ and $w$

$$y^{\mathrm{GE}} \approx \frac{\tau^{\mathrm{UEG}}}{\tau^{\mathrm{UEG}} + \frac{1}{72}\frac{|\nabla\rho|^2}{\rho}} = \frac{1}{1 + \frac{1}{72}\frac{|\nabla\rho|^2}{\rho\tau^{\mathrm{UEG}}}} = \frac{1}{1 + \frac{5}{27}s^2} \approx 1 - \frac{5}{27}s^2 \quad (15)$$

$$w^{\mathrm{GE}} = \frac{y^{\mathrm{GE}} - 1}{y^{\mathrm{GE}} + 1} \approx -\frac{5}{54}s^2 \quad (16)$$

Using eqs 2, 7, 8, 9, and 16, we obtain the gradient expansion of the local part of the M08-type exchange

$$F_{\mathrm{x}}^{\mathrm{M08-2X\ Loc}} \approx a_0 + b_0 + \left[a_0\mu^{\mathrm{PBE}} + b_0\mu_2 - \frac{5}{54}(a_1 + b_1)\right]s^2 \quad (17)$$

Note that by dropping the second term in the parentheses of eq 5, we recover the exchange functional form used for M05-2X and M06-2X, although those functionals have different numerical coefficients than are used here. In eq 6, we use $\mu_2 = 10/81$ (as mentioned above, this is also the second-order expansion coefficient for the exact exchange) and $\kappa_2 = 0.552$; both are the same as used in the SOGGA[8] functional.

**4.2. M08-Type Correlation Functional.** In the M05-2X and M06-2X correlation functionals, we treat the opposite-spin and parallel-spin correlation differently by using an ansatz of Stoll et al.[161] for the LSDA correlation energies. However, a recent study by Gori-Giorgi and Perdew[162] shows that the Stoll ansatz is inaccurate for a uniform electron gas. Therefore, we do not use the Stoll ansatz in the M08-type functional. We also avoid using the M06-2X self-correlation correction factor, $D_\sigma$, which only solves the one- or two-electron self-correlation problem and cannot solve the self-exchange and many-electron self-interaction problems.[163,164] Furthermore, a singularity in $D_\sigma$ can lead to convergence problems in the self-consistent-field iterations.[165]

The functional form of the M08-type correlation functional is given by

$$E_{\mathrm{C}}^{\mathrm{M08}} = \int d\mathbf{r}\rho\varepsilon_{\mathrm{C}}^{\mathrm{LSDA}}(r_s,\varsigma)f_3(w)d\mathbf{r} + \int d\mathbf{r}\rho H^{\mathrm{PBE}}(r_s,\varsigma,t)f_4(w)d\mathbf{r} \quad (18)$$

where $\varepsilon_{\mathrm{C}}^{\mathrm{LSDA}}(r_s,\varsigma)$ is the correlation energy per electron of the uniform electron gas limit, for which we use the parametrization of Perdew and Wang;[166] $H^{\mathrm{PBE}}(r_s,\varsigma,t)$ is the PBE[7] gradient correction for the correlation, and $f_3(w)$ and $f_4(w)$ are the kinetic-energy-density enhancement factors for correlation

$$f_3(w) = \sum_{i=0}^{11} c_i w^i \quad (19)$$

$$f_4(w) = \sum_{i=0}^{11} d_i w^i \quad (20)$$

The arguments in $\varepsilon_{\mathrm{C}}^{\mathrm{LSDA}}(r_s,\varsigma)$ and $H^{\mathrm{PBE}}(r_s,\varsigma,t)$ are defined by

$$r_{\mathrm{s}} = (3/4\pi\rho) \quad (21)$$

$$\zeta = (\rho_\alpha - \rho_\beta)/(\rho_\alpha + \rho_\beta) \quad (22)$$

$$t = |\nabla\rho|/[4(3/\pi)^{1/6}\rho^{7/6}] \quad (23)$$

**4.3. Hybrid Meta Functional.** The hybrid meta exchange-correlation energy can be written as follows

$$E_{\mathrm{xc}}^{\mathrm{hyb}} = Y \times E_{\mathrm{x}}^{\mathrm{HF}} + (1 - Y)E_{\mathrm{x}}^{\mathrm{Loc}} + E_{\mathrm{c}}^{\mathrm{DFT}} \quad (24)$$

where $E_{\mathrm{X}}^{\mathrm{HF}}$ is the nonlocal HF exchange energy, $Y$ is $X/100$, $X$ is the percentage of HF exchange in the hybrid functional, $E_{\mathrm{X}}^{\mathrm{Loc}}$ is the local DFT exchange energy, and $E_{\mathrm{C}}^{\mathrm{DFT}}$ is the local DFT correlation energy. The M08-type exchange enhancement factor can formally be written as

$$F_{\mathrm{x}}^{\mathrm{M08}} = YF_{\mathrm{x}}^{\mathrm{HF}} + (1 - Y)F_{\mathrm{x}}^{\mathrm{M08-Loc}} \quad (25)$$

where $F_{\mathrm{x}}^{\mathrm{HF}}$ is the factor implied by HF exchange.

We optimize $X$ along with the parameters in the two M08-type functionals. The optimization procedure is given in the next subsection.

For a slowly varying density, $F_{\mathrm{x}}^{\mathrm{HF}}$ has the second-order gradient expansion[167]

$$F_{\mathrm{x}}^{\mathrm{HF}} = 1 + \mu_2 s^2 + O(\nabla^2) \quad (26)$$

Combining eqs 7, 25, and 26, we obtain the gradient expansion of the M08-type exchange functionals

$$F_{\mathrm{x}}^{\mathrm{M08}} \approx [(1 - Y)(a_0 + b_0) + Y] + \left[(1 - Y)a_0\mu^{\mathrm{PBE}} + (Y + Yb_0)\mu_2 - \frac{5}{54}(1 - Y)(a_1 + b_1)\right]s^2 \quad (27)$$

**4.4. Optimization of the New Hybrid Meta-GGA.** All parameter optimizations were carried out in a self-consistent fashion. The parameters $a_i$, $b_i$, $c_i$, and $d_i$ in eqs 8, 9, 19, and 20 were determined by fitting to the data in the training set. To obtain the correct UEG limit, according to eqs 11 and 19, we enforce the following constraints in M08-HX:

$$(1-Y)(a_0+b_0) + Y = 1 \quad (28)$$

$$c_0 = 1 \quad (29)$$

For the M08-SO functional, we respect the gradient expansion to the second order in both exchange and correlation. According to eqs 18 and 27, we therefore enforce the following constraints:

$$(1 - Y)a_0\mu^{\mathrm{PBE}} + (Y + (1 - Y)b_0)\mu_2 - \frac{5}{54}(1 - Y)(a_1 + b_1) = \frac{10}{81} \approx 0.1235 \quad (30)$$

$$c_1 = 0 \quad (31)$$

$$d_0 = 1 \quad (32)$$

The constraints in eqs 29, 31, and 32 ensure that the M08-SO correlation functional is correct to the second order for slow varying density because, as discussed elsewhere,[7,168] the PBE correlation functional is correct through the second order.

We optimized the remaining parameters in M08-HX and M08-SO against accurate data to minimize a training function $F$ defined by

$$F = \mathrm{RMSEPB(MGAE109)} + \mathrm{RMSE(IP13)} + \mathrm{RMSE(EA13)} + \mathrm{RMSE(PA8)} + \mathrm{RMSE(DBH76)} + 10 \times \mathrm{RMSE(NCCE31)} + \mathrm{RMSE(ABDE4)} + \mathrm{RMSE(AE17)} + \mathrm{RMSE}(\pi\mathrm{TC13}) \quad (33)$$

where RMSEPB denotes RMSE per bond, and RMSE denotes root-mean-square error. Note that $\pi$TC13 is the union of $\pi$IE3, PA-P5/06, and PA-SB5/06; all databases are listed in Table 1. As explained in Sections 2 and 3, Table 1 also presents the references for each database and the basis sets and geometries employed in this work for each database. The optimized parameters for M08-HX and M08-SO are listed in Table 2.

Using the optimized parameters in Table 2 and eq 27, we obtain the second-order gradient expansion of the M08-HX exchange

$$F_{\mathrm{x}}^{\mathrm{M08-HX}} \approx 1 + \mu^{\mathrm{M08-HX}} s^2 = 1 + 0.2696 s^2 \quad (34)$$

## 5. Results and Discussion

We first discuss the performance of the tested functionals for the training set. In some cases we compare to other popular and high-performance functionals;[2,4,5,7−11,14,15,17−28,92,133,166,169−177] these functionals are explained in Table 3. Conventional functionals can be classified according to a ladder of ingredients, and Table 3 also indicates the rung of Jacob's ladder[87,88] to which each functional belongs. LSDA is rung 1, GGAs are rung 2, meta functionals that contains spin kinetic energy density or Laplacians of the density are rung 3 (meta GGAs), hybrid GGAs and hybrid meta functionals are rung 4, and using unoccupied orbitals puts one on rung 5. Some functionals on rung 4 contain rung-3 ingredients, and some do not. To distinguish these they are called respectively hybrid meta, denoted 4 (HM), and hybrid GGA, denoted 4 (HG).

Table 3 also contains a column called $\mu$. This is defined in eq 4 and is worked out using the same methods as in Sections 4.1 and 4.3.

In all tables after Table 3, the functionals will always be given in order of increasing mean unsigned error (MUE) for that table (as given in the last row or column of the table or in the following table). When meaningful, we also give mean signed error (MSE).

**5.1. Performance for the Training Sets.** Table 4 present the mean errors of M05-2X, M06-2X, M08-HX, M08-SO, and B3LYP for the training data for molecules. The TC-MUE defined in Table 4 is the MUE for the 160 data for main-group thermochemistry, TK-MUE is the MUE for the 76 data for thermochemical kinetics, and NC-MUE is the MUE for the 31 noncovalent data in the training set. The last row is for AMUE, which is the average of TC-MUE, TK-MUE, and NC-MUE. M06-2X gives smaller NC-MUE than M08-HX, and M08-HX gives the smaller TK-MUE. Overall the AMUE for M08-HX is just slightly better than M06-2X. Since we enforced the gradient expansion coefficients to the second order in M08-SO, M08-SO is just slightly worse than M06-2X and M08-HX. As shown by the AMUE in Table 4, M08-SO, M06-2X, and M08-2X perform much better than B3LYP for the molecular training set.

Table 5 lists the TC-MUEs, TK-MUEs, NC-MUEs, and AMUEs for 29 functionals and for HF theory and a column for the number of optimized parameters in each functional, including parameters inherited from incorporated functional forms, even if they are not re-optimized. Table 5 shows that LSDA performs better than HF theory for the TC and NC areas, but it is inferior to HF for kinetics. The SOGGA functional,[8] which has been designed for lattice constants in solid-state physics and to illustrate the results with a correct

***Table 3.*** Complete List of Functionals Used in This Article[a]

| method | rung | X | $\mu$[b] | refs |
|---|---|---|---|---|
| B1B95 | 4 (HM) | 28 | 0.2321 | 9 |
| B2PLYP | 5 | 53 | 0.1944 | 83 |
| B3LYP | 4 (HG) | 20 | 0.2222 | 4,5,14,15 |
| B3LYP* | 4 (HG) | 15 | 0.2160 | 174 |
| B3P86 | 4 (HG) | 20 | 0.2222 | 5, 14,170 |
| B3PW91 | 4 (HG) | 20 | 0.2222 | 5,14,171 |
| B88[c] | 2 | 0 | 0.2743 | 5 |
| B97−1 | 4 (HG) | 21 | 0.1654 | 172 |
| B97−2 | 4 (HG) | 21 | 0.0376 | 175 |
| B97−3 | 4 (HG) | 26.93 | 0.1044 | 19 |
| B98 | 4 (HG) | 21.98 | 0.1244 | 92 |
| BB1K | 4 (HM) | 42 | 0.2109 | 23 |
| BLYP | 2 | 0 | 0.2743 | 4,5 |
| BMK | 4 (HM) | 42 | 1.1112 | 25 |
| BP86 | 2 | 0 | 0.2743 | 5, 170 |
| HCTH | 2 | 0 | −0.1260 | 172 |
| HFLYP | 4 (HG) | 100 | 0.1235 | 176 |
| LSDA | 1 | 0 | 0.0000 | 2,166,169 |
| M05 | 4 (HM) | 28 | 0.1872 | 26 |
| M05-2X | 4 (HM) | 56 | 0.1889 | 27 |
| M06 | 4 (HM) | 27 | 0.1762 | 28 |
| M06-2X | 4 (HM) | 54 | 0.1881 | 28 |
| M06-HF | 4 (HM) | 100 | 0.0900 | 177 |
| M06-L | 3 | 0 | 0.2678 | 11 |
| M08-HX | 4 (HM) | 52.23 | 0.2696 | this work |
| M08-SO | 4 (HM) | 56.79 | 0.1235 | this work |
| MPW1B95 | 4 (HM) | 31 | 0.12(0.23) [d] | 24 |
| MPW1K | 4 (HG) | 42 | 0.12 (0.21) [d] | 20 |
| mPW1PW[e] | 4 (HG) | 25 | 0.12(0.24) [d] | 17 |
| mPW2PLYP | 5 | 55 | 0.12 (0.19)[d] | 90 |
| MPW3LYP | 4 (HG) | 20 | 0.12 (0.25) [d] | 24 |
| MPWB1K | 4 (HM) | 44 | 0.12 (0.21)[d] | 24 |
| OLYP | 2 | 0 | 0.0000 | 4,133 |
| PBE | 2 | 0 | 0.2195 | 7 |
| PBEh[f] | 4 (HG) | 25 | 0.1955 | 18 |
| PW91 | 2 | 0 | 0.12 (0.27) [d] | 171 |
| SOGGA | 2 | 0 | 0.1235 | 8 |
| TPSS | 3 | 0 | 0.1235 | 10 |
| TPSSh | 4 (HM) | 10 | 0.1235 | 22 |
| VSXC[e] | 3 | 0 | 0.0982 | 173 |
| $\tau$-HCTHh | 4 (HM) | 15 | 0.0733 | 21 |

[a] Hartree−Fock theory (which could be considered as a fourth-rung functional but here is considered to be a form of wave function theory (WFT)) is not included in this table. [b] The second-order gradient expansion coefficient for the exchange enhancement factor. [c] B88 denotes using Becke's 1988 exchange functional (the same exchange functional that is used in BLYP) with no correlation functional. [d] See ref 8. [e] Also called mPW1PW91, mPW0, and MPW25. [f] Also called PBE1PBE and PBE0. [g] Also called VS98.

***Table 4.*** Statistical Errors (kcal/mol) for the Molecular Training Data

| database | M08-HX MSE | M08-HX MUE | M06-2X MSE | M06-2X MUE | M08-SO MSE | M08-SO MUE | M05-2X MSE | M05-2X MUE | B3LYP MSE | B3LYP MUE |
|---|---|---|---|---|---|---|---|---|---|---|
| MGAE109 | −0.20 | 0.39 | −0.18 | 0.40 | −0.24 | 0.44 | −0.02 | 0.48 | −0.69 | 0.91 |
| IP13 | 3.20 | 3.37 | 1.06 | 2.54 | 3.13 | 3.53 | 1.69 | 3.54 | 3.58 | 4.72 |
| EA13 | −0.79 | 1.36 | 1.30 | 2.07 | −2.72 | 2.76 | 0.53 | 2.03 | −1.51 | 2.29 |
| PA8 | 0.23 | 1.01 | −0.19 | 1.75 | −0.52 | 1.57 | −0.25 | 1.23 | 0.18 | 1.02 |
| ABDE4 | −0.49 | 0.62 | 0.27 | 0.74 | 0.52 | 2.30 | −0.18 | 0.61 | −8.62 | 8.62 |
| $\pi$IE3 | 2.77 | 2.77 | 1.63 | 1.63 | 2.06 | 2.06 | 2.99 | 2.99 | 6.24 | 6.24 |
| PA-P5 | 0.16 | 0.46 | 0.37 | 0.66 | 1.48 | 1.48 | 2.07 | 2.07 | 5.79 | 5.79 |
| PA-SB5 | 2.50 | 2.50 | 1.69 | 2.00 | 1.47 | 1.88 | 3.90 | 3.90 | 5.90 | 5.90 |
| TC-MUE[a] | | 0.86 | | 0.86 | | 1.09 | | 1.10 | | 1.94 |
| HTBH38[b] | 0.00 | 0.73 | −0.51 | 1.13 | −0.51 | 1.09 | −0.39 | 1.34 | −4.13 | 4.23 |
| HATBH12[b] | −0.96 | 1.72 | −0.81 | 1.61 | −1.30 | 1.84 | 1.15 | 2.00 | −8.49 | 8.49 |
| NSBH16[b] | 0.63 | 1.10 | 0.77 | 1.22 | 0.23 | 1.06 | −0.79 | 1.48 | −3.25 | 3.25 |
| UABH10[b] | 0.39 | 1.00 | 0.32 | 0.92 | 0.27 | 1.15 | 0.91 | 1.77 | −1.42 | 2.02 |
| TK-MUE[c] | | 1.00 | | 1.20 | | 1.21 | | 1.53 | | 4.40 |
| HB6[d] | −0.03 | 0.31 | −0.14 | 0.25 | 0.06 | 0.23 | −0.05 | 0.20 | −0.93 | 0.93 |
| CT7[d] | 0.06 | 0.32 | −0.01 | 0.27 | 0.19 | 0.50 | 0.13 | 0.30 | 0.30 | 0.54 |
| DI6[d] | −0.01 | 0.28 | −0.12 | 0.31 | 0.09 | 0.20 | −0.15 | 0.32 | −0.94 | 0.94 |
| WI7[d] | −0.02 | 0.09 | 0.06 | 0.09 | −0.03 | 0.05 | 0.00 | 0.03 | −0.39 | 0.39 |
| PPS5[d] | −0.42 | 0.45 | −0.33 | 0.39 | −0.30 | 0.43 | −0.69 | 0.71 | −3.19 | 3.19 |
| NC-MUE[e] | | 0.28 | | 0.25 | | 0.28 | | 0.29 | | 1.09 |
| AMUE | | 0.71 | | 0.77 | | 0.86 | | 0.97 | | 2.48 |

[a] This is the MUE for MGTC160. [b] HTBH38, HATBH12, NSBH16, and UABH10 are components of DBH76 as explained in refs 11 and 28. [c] This is the MUE for DBH76. [d] HB6, CT7, DI6, WI7, and PPS5 are components of NCCE31, as explained in refs 28 and 113. [e] This is the MUE for NCCE31.

***Table 5.*** Statistical Errors (kcal/mol) for the Molecular Training Data

| functionals | rung | no. of parameters[a] | TC-MUE | TK-MUE | NC-MUE | AMUE |
|---|---|---|---|---|---|---|
| M08-HX | 4 (HM) | 47 | 0.86 | 1.00 | 0.28 | 0.71 |
| M06-2X | 4 (HM) | 35 | 0.86 | 1.20 | 0.25 | 0.77 |
| M08-SO | 4 (HM) | 44 | 1.09 | 1.21 | 0.28 | 0.86 |
| M05-2X | 4 (HM) | 22 | 1.10 | 1.53 | 0.29 | 0.97 |
| BMK | 4 (HM) | 20 | 1.25 | 1.29 | 1.12 | 1.22 |
| MPWB1K | 4 (HM) | 7 | 1.74 | 1.37 | 0.65 | 1.25 |
| M06 | 4 (HM) | 38 | 1.28 | 2.13 | 0.41 | 1.28 |
| M06-HF | 4 (HM) | 38 | 1.33 | 2.25 | 0.42 | 1.33 |
| M05 | 4 (HM) | 22 | 1.54 | 2.03 | 0.44 | 1.34 |
| BB1K | 4 (HM) | 6 | 2.07 | 1.29 | 1.18 | 1.51 |
| MPW1B95 | 4 (HM) | 7 | 1.37 | 2.66 | 0.67 | 1.57 |
| B97−3 | 4 (HG) | 19 | 1.69 | 1.87 | 1.19 | 1.58 |
| B1B95 | 4 (HM) | 6 | 1.40 | 2.53 | 1.26 | 1.73 |
| MPW1K | 4 (HG) | 5 | 3.07 | 1.55 | 0.87 | 1.83 |
| B97−2 | 4 (HG) | 16 | 1.77 | 2.74 | 1.22 | 1.91 |
| B98 | 4 (HG) | 16 | 1.61 | 3.78 | 0.73 | 2.04 |
| mPW1PW | 4 (HG) | 5 | 1.91 | 3.39 | 0.89 | 2.07 |
| B97−1 | 4 (HG) | 16 | 1.63 | 3.93 | 0.66 | 2.07 |
| PBEh | 4 (HG) | 1 | 1.78 | 3.87 | 0.67 | 2.11 |
| M06-L | 3 | 39 | 1.92 | 4.02 | 0.58 | 2.17 |
| B3LYP | 4 (HG) | 7 | 1.94 | 4.40 | 1.09 | 2.48 |
| MPW3LYP | 4 (HG) | 8 | 1.83 | 5.02 | 0.80 | 2.55 |
| $\tau$-HCTHh | 4 (HM) | 20 | 1.78 | 4.92 | 1.47 | 2.73 |
| TPSSh | 4 (HM) | 1 | 2.18 | 6.44 | 1.03 | 3.21 |
| TPSS | 3 | 0 | 2.17 | 8.33 | 1.19 | 3.90 |
| BLYP | 2 | 4 | 2.45 | 8.31 | 1.58 | 4.11 |
| PBE | 2 | 0 | 3.15 | 8.92 | 1.10 | 4.39 |
| HFLYP | 4 (HG) | 3 | 7.84 | 7.08 | 0.76 | 5.22 |
| SOGGA | 2 | 0 | 6.31 | 11.45 | 2.12 | 6.63 |
| LSDA | 1 | 0 | 13.45 | 15.13 | 2.17 | 10.25 |
| HF | NA[a] | 0 | 26.48 | 7.71 | 2.62 | 12.27 |

[a] Number of independent optimized parameters in each functional. This does not include hidden parameters (such as the choice of a functional form that makes a term in the gradient expansion vanish for empirical reasons) or parameters that are fitted to accurate calculations of the correlation energy of a uniform electron gas. [b] NA denotes "not applicable".

second-order gradient expansion, performs better than LSDA for all three areas, but it is not as good for main-group

***Table 6.*** Statistical Errors (kcal/mol) for Atomic Training Data (AE17)

| method | X[a] | MSE | MUE |
|---|---|---|---|
| Results with the aug-cc-pVQZ Basis Set | | | |
| M06-2X | 54 | 0.61 | 2.00 |
| M08-HX | 52.23 | −0.70 | 4.05 |
| M06 | 27 | 0.33 | 4.37 |
| B98 | 21.98 | 3.08 | 5.02 |
| B97−1 | 21 | 2.03 | 5.38 |
| M06-L | 0 | −5.25 | 5.61 |
| M08-SO | 56.79 | 4.82 | 5.71 |
| $\tau$-HCTHh | 15 | 0.69 | 6.04 |
| M06-HF | 100 | −5.58 | 6.20 |
| B97−3 | 26.93 | 2.83 | 6.61 |
| M05-2X | 56 | −7.88 | 7.89 |
| HFLYP | 100 | −7.61 | 8.37 |
| MPW1K | 42 | −8.52 | 9.22 |
| B97−2 | 21 | −8.74 | 9.69 |
| mPW1PW | 25 | −8.82 | 9.77 |
| M05 | 28 | −7.04 | 9.98 |
| BLYP | 0 | −9.29 | 10.05 |
| TPSSh | 10 | −14.32 | 14.32 |
| BB1K | 42 | −13.45 | 14.74 |
| B1B95 | 28 | −13.73 | 15.19 |
| MPWB1K | 44 | −13.44 | 15.32 |
| MPW1B95 | 31 | −13.69 | 15.94 |
| B3LYP | 20 | −16.90 | 16.90 |
| TPSS | 0 | −16.94 | 16.94 |
| MPW3LYP | 20 | −16.98 | 16.98 |
| BMK | 42 | 17.77 | 18.24 |
| PBEh | 25 | 39.48 | 39.57 |
| PBE | 0 | 48.81 | 48.81 |
| B88 | 0 | 190.46 | 190.46 |
| HF | 100 | 191.97 | 191.97 |
| SOGGA | 0 | 284.46 | 284.46 |
| LSDA | 0 | 425.54 | 425.54 |
| Results with Other Basis Sets | | | |
| M06-2X/MQZVP | 54 | −0.73 | 1.76 |
| M06-2X/MG3S | 54 | 4.38 | 4.52 |
| M05-2X/MG3S | 56 | −3.86 | 4.83 |
| M05-2X/MQZVP | 56 | −9.27 | 9.29 |

[a] Percentage of Hartree−Fock exchange in each functional.

***Table 7.*** Statistical Errors (kcal/mol) for the G3-3AE75 Database[a]

| methods | rung | MaxE+ | MaxE- | MSE | MUE |
|---|---|---|---|---|---|
| | | with Spin−Orbit | | | |
| M08-HX/MG3SXP | 4 (HM) | 8.2 ($C_4H_4N_2$) | −8.6 ($SO_3$) | −0.54 | 2.28 |
| M08-HX/6−311++G(3df,3pd) | 4 (HM) | 8.7 ($C_4H_4N_2$) | −12.5 ($SO_3$) | 0.11 | 2.54 |
| M06-2X/6−311++G(3df,3pd) | 4 (HM) | 17.8 ($P_4$) | −9.9 ($SO_3$) | 0.70 | 2.72 |
| M06-SO/MG3SXP | 4 (HM) | 8.9 ($C_4H_4N_2$) | −12.4 ($SF_6$) | −1.41 | 2.76 |
| M06-2X/MG3SXP | 4 (HM) | 15.5 ($P_4$) | −6.7 ($SO_3$) | −0.38 | 2.86 |
| M05-2X/MG3SXP | 4 (HM) | 16.6 ($P_4$) | −8.2 ($SO_3$) | 3.47 | 4.11 |
| TPSS/6−311++G(3df,3pd)[b] | 4 (HM) | 12.3 ($C_4H_4N_2$) | −9.4 ($PF_5$) | 4.19 | 4.76 |
| B3LYP/6−311+G(3df,3pd)[b] | 4 (HG) | 4.5 ($C_4H_4N_2$) | −23.7 ($SF_6$) | −9.23 | 9.39 |
| | | without Spin−Orbit | | | |
| M08-HX/MG3SXP | 4 (HM) | 8.5 ($C_4H_4N_2$) | −7.3 ($SO_3$) | 0.46 | 2.31 |
| M08-HX/6−311++G(3df,3pd) | 4 (HM) | 9.0 ($C_4H_4N_2$) | −11.2 ($SO_3$) | 1.10 | 2.79 |
| M08-SO/MG3SXP | 4 (HM) | 9.2 ($C_4H_4N_2$) | −9.6 ($SF_6$) | −0.42 | 2.68 |
| M06-2X/MG3SXP | 4 (HM) | 15.5 ($P_4$) | −5.8 ($Si(CH_3)_4$) | 0.61 | 2.95 |
| M06-2X/6−311++G(3df,3pd) | 4 (HM) | 17.8 ($P_4$) | −8.7 ($SO_3$) | 1.70 | 3.06 |
| M05-2X/MG3SXP | 4 (HM) | 16.6 ($P_4$) | −6.9($SO_3$) | 4.46 | 4.90 |
| | | Results from Literature (without Spin−Orbit) | | | |
| Mpw2plyp-D[d] | 5 | 5.9 ($C_4H_4N_2$) | −9.1 ($P_4$) | −0.40 | 2.11 |
| B2PLYP-D[d] | 5 | 6.8 ($C_4H_4N_2$) | −8.7 ($Si(CH_3)_4$) | −0.74 | 2.21 |
| TPSSh[c] | 4 (HM) | 6.6 ($C_8H_{18}$) | −16.2 (PF5) | −0.16 | 3.33 |
| mPW2PLYP[d] | 5 | 4.8 ($C_4H_4N_2$) | −9.6 ($Si(CH_3)_4$) | −2.96 | 3.39 |
| B2PLYP[d] | 5 | 5.2 ($C_4H_4N_2$) | −13.6 ($Si(CH_3)_4$) | −4.27 | 4.67 |
| VSXC[c] | 3 | 8.7 ($C_6H_5$) | −12.0 ($C_8H_{18}$) | −1.97 | 4.74 |
| B3PW91[c] | 4 (HG) | 17.0 (naphthalene) | −17.0 ($PF_5$) | 2.54 | 4.87 |
| LC-$\omega$PBE[e] | 4 | N. A.[f] | N. A.[f] | 2.05 | 5.28 |
| TPSS[c] | 3 | 12.8 ($S_2Cl_2$) | −7.5 ($PF_5$) | 5.19 | 5.48 |
| OLYP[c] | 2 | 11.0 ($CF_3$) | −20.9 ($Si(CH_3)_4$) | −6.41 | 7.91 |
| B3LYP[c] | 4 (HG) | 4.9 ($C_4H_4N_2$) | −20.8 ($SF_6$) | −8.23 | 8.44 |
| HCTH[c] | 2 | 22.2 ($C_2F_6$) | −27.5 ($Si(CH_3)_4$) | −6.38 | 10.18 |
| PBE0[c] | 4 (HG) | 35.6 (naphthalene) | −14.5 ($PF_5$) | 9.28 | 10.20 |
| BPW91[c] | 2 | 28.0 (azulene) | −22.4 ($Si(CH_3)_4$) | 4.97 | 11.08 |
| PKZB[c] | 3 | 11.0 ($P_4$) | −35.4 ($PF_5$) | −10.59 | 11.24 |
| BLYP[c] | 2 | 11.0 ($C_4H_4N_2$) | −41.0 ($C_8H_{18}$) | −12.42 | 13.88 |
| PBE[c] | 2 | 79.7 (azulene) | none | 32.77 | 32.77 |
| PW91[c] | 2 | 81.1 (azulene) | none | 35.25 | 35.25 |
| BP86[c] | 2 | 72.7 (azulene) | none | 38.61 | 38.61 |
| B3P86[c] | 3 | 79.2 ($C_8H_{18}$) | none | 41.89 | 41.89 |
| LSDA[c] | 1 | 347.5 (azulene) | none | 197.11 | 197.1 |
| HF[c] | NA | | −582.2 ($C_8H_{18}$) | −336.4 | 336.4 |

[a] B3LYP/6−31G(2df,p) geometries are used. [b] Calculated from the raw energies in the Supporting Information of ref 22. [c] Taken from ref 22. [d] Calculated from the results in the Supporting Information of ref 90. [e] Calculated from the results in ref 70. [f] N. A. denotes "not available". The maximum errors for LC-$\omega$PBE were not reported in ref 70.

energetics as PBE and BLYP, which are the most popular functionals on rung 2. The M06-L meta GGA, a third-rung functional, outperforms the most popular hybrid GGA, which is the fourth-rung B3LYP functional. Table 5 shows that B97−3 is the best performing hybrid GGA for the molecular training set, and BMK is the best performing non-Minnesota functional.

With 13 more semiempirical parameters, the AMUE of M06-2X is 0.2 kcal/mol smaller than that of M05-2X, whereas it is just 0.06 kcal/mol greater than the AMUE of M08-HX with 12 less semiempirical parameters. Furthermore, M06-2X has 9 less parameters than the M08-SO functional, but M06-2X outperforms M08-SO by a small margin. Tables 4 and 5 indicate that an AMUE of ∼0.7 kcal/mol is the limit of accuracy of the hybrid meta functionals for this set of 267 molecular data.

Table 6 lists the mean errors for the atomic training data. As shown in this table, M06-2X performs better than M08-HX and M08-SO for atomic energies. The results do not correlate with the percentage of Hartree−Fock exchange.

**5.2. Performance for the Test Sets.** In this section, we present tests against some databases which are outside of our training set.

*5.2.1. G3-3AE75 Database.* Table 7 presents the mean signed errors (MSEs) and mean unsigned errors (MUEs) for the molecules in the G3-3 database. The G3-3 data set[33] contains molecules as large as naphthalene and multihalogen-containing molecules such as $SF_6$ and $PF_5$. The tests for this database in the literature employed the 6−311++G(3df,3pd) basis set without spin−orbit energies. To make a consistent comparison, we calculated the MUEs and MSEs with and without spin−orbit energies for the M06-2X, M08-HX, TPSS, and B3LYP functionals.

Table 7 shows that including spin−orbit energies improves the performance of M06-2X and M08-HX by 0.1−0.3 kcal/mol but deteriorates the performance of M08-SO by ∼0.1 kcal/mol. (Nevertheless spin−orbit coupling is a real effect, and it should always be included. Except for Table 6 and the bottom section of Table 7, all other results in this paper and in our previous work (except when comparing to theoretical nonrelativistic data for atoms) include spin−orbit

***Table 8.*** Results for the LMAE14 Database (kcal/mol)

| molecule | exp. | M08-HX | M08-SO | M06-2X | M05-2X | B3LYP |
|---|---|---|---|---|---|---|
| $C_6F_6$ | 1389.9 | 1398.2 | 1397.4 | 1406.4 | 1410.1 | 1383.7 |
| $C_6F_5Cl$ | 1365.1 | 1373.0 | 1371.8 | 1379.7 | 1383.1 | 1353.5 |
| dodecane $C_{12}H_{26}$ | 3655.4 | 3651.1 | 3647.6 | 3648.0 | 3659.3 | 3622.9 |
| hexadecane $C_{16}H_{34}$ | 4833.7 | 4827.3 | 4822.7 | 4823.1 | 4837.9 | 4786.1 |
| adamantane $C_{10}H_{16}$ | 2695.1 | 2693.7 | 2691.8 | 2690.3 | 2702.9 | 2657.7 |
| diadamantane $C_{14}H_{20}$ | 3624.9 | 3624.6 | 3622.3 | 3620.1 | 3639.1 | 3568.2 |
| pyrene $C_{16}H_{10}$ | 3301.9 | 3298.6 | 3294.8 | 3302.8 | 3319.6 | 3273.7 |
| fluroanthene $C_{16}H_{10}$ | 3286.7 | 3284.3 | 3281.0 | 3288.1 | 3304.5 | 3259.5 |
| anthracene $C_{14}H_{10}$ | 2952.8 | 2949.9 | 2946.8 | 2953.5 | 2967.8 | 2930.9 |
| phenazine $C_{12}H_8N_2$ | 2693.2 | 2700.2 | 2698.0 | 2698.6 | 2711.3 | 2683.5 |
| azobenzene $C_{12}H_{10}N_2$ | 2793.6 | 2797.5 | 2796.2 | 2795.4 | 2807.7 | 2782.1 |
| benzophenone $C_{13}H_{10}O$ | 2882.3 | 2881.8 | 2879.5 | 2884.8 | 2896.4 | 2861.5 |
| dibenzothiophene $C_{12}H_8S$ | 2562.7 | 2561.1 | 2558.2 | 2562.7 | 2576.0 | 2536.5 |
| dithiin $C_{16}H_{12}S_2$ | 3503.1 | 3510.4 | 3508.1 | 3511.5 | 3529.1 | 3472.9 |
| MSE | | 0.8 | −1.7 | 1.7 | 14.6 | −26.3 |
| MUE | | 4.1 | 5.5 | 5.7 | 14.6 | 26.3 |

[a] M06-L/6−311+G(2df,2p) geometries are used.

energy for cases where it is nonzero.) M08-HX performs better than M06-2X especially when using the MG3SXP basis. M06-2X, M08-SO, and M08-2X outperform the B2-PLYP and mPW2-PLYP functionals, but they underperform the B2-PLYP-D and mPW2PLYP-D functionals. The ability of the fourth-rung M06-2X, M08-SO, and M08-2X functionals to compete with fifth-rung functionals is very encouraging.

*5.2.2. LMAE14 Database.* Table 8 present the results for the LMAE14 database. This is a data set of large molecules, and B3LYP gives an error of 26.2 kcal/mol for this database, whereas M08-HX gives an MUE of only 4.1 kcal/mol. M06-2X, with an MUE of 5.7 kcal/mol, is less accurate than M08-SO and M08-HX. O3LYP is the best functional for this database in the test of Curtiss et al.,[118] with an MUE of 8.6 kcal/mol, but this is larger than that for M08-HX by more than a factor of 2; this illustrates the tremendous progress that has been made in the last four years. The results for this database show that M08-HX does improve upon M06-2X for large molecules.

*5.2.3. IE34 Database.* Tables S2 and 9 present the results for the isomerization energy database. The changes in structure and bonding for the 34 isomerizations are notably diverse and potentially challenging. B3LYP gives a large error for octane isomerization (Table S1, entry 11), and Grimme[39,44] also pointed out the inability of most popular functionals to describe this type of stereoelectronic effect. M06-2X, M08-SO, and M08-HX perform well for describing the stereoelectronic effects in hydrocarbons, as does M05-2X.[94] M06-2X performs poorly for reactions involving three-member-ring molecules (Table S1, entries 3, 8, 16, and 25), and two M08 functionals perform better in these cases.

The statistical errors for IE34 are given in Table 9. We also include the results for the best performing GGA, hybrid GGA, meta-GGA, and doubly hybrid functional and a WFT method in the test of Grimme et al.[45] Table 9 shows that M06-HX gives a smaller MUE and RMSE than M06-2X. The number of outliers for the two M08 functionals is also less than that for M06-2X.

Table 9 also shows that M06-2X and two M08 functionals are more accurate than mPW2-PLYP for the IE34 database, but they are less accurate than SCS-MP2, the best performing method in Table 10, which is a WFT method for which the computation for an $N$-atom system scales as $N^5$ whereas the conventional algorithm for hybrid functional scales as $N^4$, and density fitting algorithm for rungs 1−3 have a scaling of $N^3$.

***Table 9.*** Statistical Errors for the IE34 Database

| method | rung | MaxE | no. outliers[a] | RMSE | MUE |
|---|---|---|---|---|---|
| M08-HX | 4 (HM) | 3.1 (33) | 2 | 1.44 | 1.12 |
| M08-SO | 4 (HM) | 3.2 (16) | 1 | 1.52 | 1.15 |
| M06-2X | 4 (HM) | 4.3 (8) | 4 | 1.65 | 1.15 |
| M05-2X | 4 (HM) | 4.6 (27) | 2 | 1.74 | 1.31 |
| B3LYP | 4 (HG) | 10.1 (12) | 9 | 3.22 | 2.28 |
| Results from Literature | | | | | |
| SCS-MP2[b,c] | WFT | 2.6 (2) | 0 | 1.27 | 1.03 |
| mPW2PLYP[b] | 5 | 6.1 (12) | 4 | 1.83 | 1.19 |
| BMK[b] | 4 (HM) | 4.7 (7) | 4 | 1.79 | 1.28 |
| PBE0[b] | 4 (HG) | 7.0 (11) | 7 | 2.45 | 1.79 |
| PBE[b] | 2 | 7.3 (11) | 6 | 2.54 | 1.89 |
| TPSS[b] | 3 | 11.4 (27) | 10 | 3.46 | 2.52 |

[a] Number of unsigned errors >3.0 kcal/mol. [b] Taken from ref 45. [c] WFT: wave function theory (not DFT).

*5.2.4. Noncovalent Databases.* The ability of the new generation of hybrid meta density functionals[13,24,27,28,113] to treat noncovalent interaction energies that are dominated by medium-range correlation is a major step forward in the usefulness of DFT for practical simulation on biological systems and soft materials. Table 10 shows, for example, that two M08 functionals and M06-2X reduce the error for the "dispersion dominated" complexes of the S22 database by an order of magnitude, as compared to the popular B3LYP functional. The reduction in error is also significant for the hydrogen bonded ones. In Tables S2 and S3 (Supporting Information), we present results for a double-$\zeta$ basis set. We also defined a quantity called mean averaged MUE (MAMUE) in Tables S2 and S3 which is an average over all three types of interactions and over CP-corrected and uncorrected results. Table S3 presents the MAMUEs for 34 functionals. If we use MAMUE to rank these functionals, we can see that the best performing GGA is SOGGA, the best performing meta-GGA is M06-L, and the best performing hybrid meta GGAs are M06-2X, M08-SO, and M08-HX.

***Table 10.*** Results (kcal/mol) for the S22 Database

| complex | best estimate | MP2/CBS | M06-2X CP | M06-2X noCP | M08-SO CP | M08-SO noCP | M08-HX CP | M08-HX noCP | B3LYP CP | B3LYP noCP |
|---|---|---|---|---|---|---|---|---|---|---|
| Hydrogen Bonded (HB) Complexes | | | | | | | | | | |
| $(NH_3)_2$ | −3.2 | −3.2 | −3.2 | −3.3 | −3.2 | −3.3 | −3.4 | −3.6 | −2.2 | −2.3 |
| $(H_2O)_2$ | −5.0 | −5.0 | −5.1 | −5.4 | −5.1 | −5.4 | −5.1 | −5.6 | −4.5 | −4.8 |
| formic acid dimmer | −18.6 | −18.6 | −18.9 | −19.6 | −18.4 | −18.9 | −18.4 | −19.3 | −17.2 | −17.8 |
| formamide dimmer | −16.0 | −15.9 | −15.6 | −16.0 | −15.6 | −16.0 | −15.8 | −16.4 | −13.9 | −14.3 |
| uracil dimer | −20.7 | −20.6 | −19.4 | −19.8 | −19.4 | −19.8 | −19.7 | −20.2 | −17.8 | −18.1 |
| 2-pyridoxine·2-aminopyridine | −16.7 | −17.4 | −15.5 | −15.8 | −15.1 | −15.5 | −15.5 | −16.0 | −13.7 | −14.0 |
| adenine·thymine WC | −16.4 | −16.5 | −15.0 | −15.3 | −14.5 | −15.0 | −15.0 | −15.5 | −12.8 | −13.1 |
| MSE-HB | | −0.1 | 0.5 | 0.2 | 0.7 | 0.4 | 0.5 | 0.0 | 1.9 | 1.3 |
| MUE-HB | | 0.1 | 0.7 | 0.6 | 0.8 | 0.6 | 0.6 | 0.6 | 2.0 | 1.6 |
| Dispersion Dominated (DD) Complexes | | | | | | | | | | |
| $(CH_4)_2$ | −0.5 | −0.5 | −0.4 | −0.5 | −0.5 | −0.5 | −0.3 | −0.3 | 0.4 | 0.4 |
| $(C_2H_4)_2$ | −1.5 | −1.6 | −1.6 | −1.7 | −1.8 | −2.0 | −1.9 | −2.0 | 0.5 | 0.5 |
| benzene·$CH_4$ | −1.5 | −1.9 | −1.5 | −1.7 | −1.6 | −2.1 | −1.8 | −2.3 | 0.8 | 0.6 |
| benzene dimer | −2.7 | −5.0 | −2.8 | −3.7 | −3.0 | −4.2 | −3.0 | −4.5 | 3.8 | 3.0 |
| pyrazine dimer | −4.4 | −6.9 | −4.2 | −4.8 | −4.2 | −5.0 | −4.2 | −5.2 | 2.6 | 2.0 |
| uracil dimer | −10.1 | −11.4 | −9.9 | −11.1 | −9.3 | −10.7 | −9.3 | −11.2 | −0.9 | −1.9 |
| indole·benzene | −5.2 | −8.1 | −4.6 | −5.7 | −4.8 | −6.2 | −4.8 | −6.6 | 4.8 | 3.9 |
| adenine·thymine stack | −12.2 | −14.9 | −12.2 | −13.4 | −11.8 | −13.2 | −11.8 | −13.7 | 1.5 | 0.5 |
| MSE | | −1.5 | 0.1 | −0.6 | 0.2 | −0.7 | 0.2 | −1.0 | 6.5 | 5.9 |
| MUE | | 1.5 | 0.2 | 0.6 | 0.3 | 0.7 | 0.4 | 1.0 | 6.5 | 5.9 |
| Mixed Complexes | | | | | | | | | | |
| ethene·ethyne | −1.5 | −1.7 | −1.4 | −1.5 | −1.5 | −1.7 | −1.6 | −1.9 | −0.6 | −0.8 |
| benzene·$H_2O$ | −3.3 | −3.6 | −3.6 | −4.1 | −3.7 | −4.2 | −3.7 | −4.4 | −1.2 | −1.7 |
| benzene·$NH_3$ | −2.4 | −2.7 | −2.4 | −2.7 | −2.5 | −2.9 | −2.7 | −3.1 | −0.1 | −0.4 |
| benzene· HCN | −4.5 | −5.2 | −4.9 | −5.4 | −5.3 | −5.9 | −5.4 | −6.2 | −2.0 | −2.4 |
| benzene dimer | −2.7 | −3.6 | −2.4 | −2.9 | −2.5 | −3.2 | −2.4 | −3.3 | 1.0 | 0.5 |
| indole·benzene T-shape | −5.7 | −7.0 | −5.1 | −5.8 | −5.2 | −6.1 | −5.2 | −6.3 | −0.5 | −1.1 |
| phenol dimer | −7.1 | −7.8 | −6.6 | −7.1 | −6.6 | −7.1 | −6.6 | −7.3 | −2.9 | −3.4 |
| MSE-mixed | −0.6 | 0.1 | −0.3 | 0.0 | −0.6 | −0.1 | −0.8 | 3.0 | 2.6 | |
| MUE-mixed | | 0.6 | 0.4 | 0.3 | 0.4 | 0.6 | 0.4 | 0.8 | 3.0 | 2.6 |
| AMUE[b] | | 0.8 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.8 | 3.8 | 3.3 |

[a] The 6−311+G(3df,2p) basis set is used for all calculations. [b] Average of the MUEs for three types of noncovalent complexes.

***Table 11.*** Interaction Energies (kcal/mol) in the UUU7 Database[a]

| | best estimate[b] | M08-SO | M06-2X | M08-HX | M05-2X | B3LYP |
|---|---|---|---|---|---|---|
| HB/HB 1 | −37.8 | −37.78 | −37.80 | −38.68 | −38.11 | −34.70 |
| HB/HB 2 | −37.4 | −36.86 | −36.49 | −37.44 | −36.81 | −32.31 |
| T/T 1 | −36.6 | −36.02 | −36.20 | −36.60 | −35.82 | −25.80 |
| HB/HB 3 | −33.5 | −33.38 | −33.10 | −33.97 | −33.46 | −30.08 |
| S/T 1 | −33.2 | −32.96 | −33.34 | −34.01 | −30.98 | −17.68 |
| HB/S 1 | −32.8 | −32.83 | −33.38 | −33.98 | −31.01 | −17.87 |
| S/S 1 | −20.1 | −19.94 | −20.51 | −21.00 | −17.06 | 1.15 |
| MSE | | 0.23 | 0.08 | −0.61 | 1.08 | 10.59 |
| MUE | | 0.24 | 0.40 | 0.61 | 1.37 | 10.59 |

[a] See Figure 1 for the structures for the peptides. [b] Calculated from the results in ref 122.

Table 11 presents results for interaction energies in large uracil trimers. Table 11 shows that the improvements for the uracil trimer database are even greater than those for S22. In particular, M08-HX, M06-2X, and M08-SO have MUEs respectively 17, 26, and 44 times lower than B3LYP.

*5.2.5. APCE5 Database.* Table 12 compares four density functionals and one DFT-D method for the aromatic peptide conformational energy database. The mean unsigned error for M05-2X, M08-HX, and M08-SO is only 22% higher than that for the functional with explicit dispersion corrections (TPSS-D), and it is four times smaller than B3LYP's mean unsigned error.

*5.2.6. BBH7/08 Database.* One goal of our development efforts is to design a density functional with accurate performance for a broad range of observables, and the prediction of barrier heights is a central concern. These reactions are chosen to include singlets, doublets, and triplets and some highly correlated systems, like $C_2H$, so they provide a significant challenge. The performance for the new benchmark barrier heights in Table 13 is very encouraging, especially in that the M06-2X and M08-2X functionals are more accurate than B3LYP for all the reactions that involve the multireference $C_2H$ molecule. These three functionals have $X = 54$, 54, and 20, respectively. Although any functional with $X$ as large as 20 (or larger) is not expected to be reliable for multireference systems, it is encouraging that the M06-2X and M08-HX give useful accuracy in these difficult cases. Further study of cases with multireference character is a worthwhile goal.[178]

The mean unsigned error of M06-2X, 1.06 kcal/mol, is better than the value, 1.20 kcal/mol, obtained for DBH76 (Table 4), and the mean unsigned error for M08-HX, 0.86 kcal/mol, is outstanding again improving on the value (1.0 kcal/mol in Table 4) for DBH76.

*5.2.7. Internuclear Distances.* The hybrid meta functionals M05-2X, M06-2X, M08-SO, and M08-2X have a high Hartree−Fock exchange (the percentage, $X$, of Hartree−Fock exchange is greater than 50%), and, as a consequence, they do not improve on B3LYP for bond lengths in small molecules, as shown in Table S4 (in the Supporting Information). Among the three high-$X$ functionals, M08-HX and

***Table 12.*** Conformational Energies (kcal/mol) in the APCE5 Database[a]

| | best estimate[b] | TPSS-D[b] | M05-2X | M08-HX | M08-SO | M06-2X | TPSS[b] | B3LYP |
|---|---|---|---|---|---|---|---|---|
| WG_12-WG_01 | 2.45 | 2.41 | 2.28 | 1.90 | 1.54 | 1.70 | 4.96 | 5.22 |
| WGG_13-WGG_01 | 4.24 | 5.35 | 3.79 | 2.58 | 2.81 | 2.48 | 9.00 | 9.32 |
| FGG_300-FGG_099 | 3.12 | 1.49 | 1.01 | 3.27 | 2.99 | 2.72 | −3.35 | −3.82 |
| GGF_01-GGF_15 | 2.93 | 2.89 | 1.82 | 1.62 | 1.48 | 1.50 | 4.00 | 3.88 |
| GFA_16-GFA_15 | 1.56 | 1.22 | 1.66 | 1.83 | 1.79 | 1.55 | 0.79 | 1.56 |
| MSE | | −0.19 | −0.75 | −0.62 | −0.74 | −0.87 | 0.22 | 0.37 |
| MUE | | 0.63 | 0.79 | 0.79 | 0.83 | 0.87 | 3.12 | 3.15 |

[a] See Figure 2 for the structures for the peptides. [b] Calculated from the results in the Supporting Information of ref 121.

***Table 13.*** Barrier Heights (kcal/mol) in the BBH7 Database

| reaction | best estimate | M08-HX | M06-2X | M08-SO | M05-2X | B3LYP |
|---|---|---|---|---|---|---|
| $CH_3 + CH_4 \rightarrow CH_4 + CH_3$ | 17.82 | 17.47 | 16.80 | 17.47 | 16.88 | 15.65 |
| $HCC + HCCH \rightarrow HCCH + CCH$ | 12.79 | 12.27 | 13.15 | 11.20 | 10.27 | 9.13 |
| $OH^- + CH_3OH \rightarrow CH_3OH + OH^-$ | 14.40 | 14.23 | 14.17 | 15.54 | 12.66 | 11.37 |
| $HCC + H_2 \rightarrow HCCH + H$ | 2.07 | 1.27 | 1.93 | 0.25 | 0.03 | 0.12 |
| $HCCH + H \rightarrow HCC + H_2$ | 32.32 | 34.49 | 31.67 | 35.25 | 35.04 | 30.33 |
| $O + CH_4 \rightarrow OH + CH_3$ | 14.19 | 13.04 | 11.80 | 13.53 | 10.95 | 7.21 |
| $OH + CH_3 \rightarrow O + CH_4$ | 9.12 | 8.26 | 6.36 | 8.63 | 5.47 | 4.48 |
| MSE | | −0.24 | −0.98 | −0.12 | −1.63 | −3.49 |
| MUE | | 0.86 | 1.08 | 1.28 | 2.41 | 3.49 |

M08-SO perform better than M05-2X and M06-2X for bond lengths in small molecules.

Both medium-range correlation energies and repulsive interactions play important roles in large crowded molecules, such as cyclophanes,[123,179] and so they provide challenging tests of density functionals. For example, B3LYP gives large errors[124] for internuclear distances in cyclophanes.[180] Table S5 compares the calculated internuclear distances in 2,6,15-trithia-*in*-[3,[4,10]][7]metacyclophane to the experimental results of Pascal et al.[123] The results for PBE, B3LYP, and MP2 were taken from Grimme.[124] Table S5 shows that B3LYP overestimates the internuclear distances in this cyclophane, whereas the popular WFT method MP2 underestimates them. Both M08-HX and M06-2X outperform B3LYP and MP2 by a large margin for the prediction of internuclear distances in this cyclophane. Table S6 compares the calculated internuclear distances in [2,2]paracyclophane to the experimental results.

Table S6 shows the same trends as in Table S5. Note that the M06-L and PBEh functionals performs fairly well for both cylcophanes.

It is instructive to average the errors over some of the key distances in the cyclophanes, in particular we consider C1−C2, C3−C14, and C4−C13 of Figure 4 and C4−C8, C4−C9, and H(C4)-C9 of Figure 3. The mean signed errors in Å are −0.03 Å for MP2, +0.05 Å for B3LYP, +0.03 Å for PBE and TPSS, +0.007 Å for M06-L, +0.006 Å for M06-2X, −0.005 Å for M08-HX, and −0.004 Å for M08-SO. The excellent performance of the new functionals is especially striking since Grimme[124] had concluded that "an explicit account of dispersive-type electron correlation effects between the clamped aromatic units is essentially for a quantitative description of cylcophane structures". The new functionals contain dispersion-like and steric exchange repulsion effects implicitly rather than as explicit molecular mechanics additions.

Table 14 averages the errors in bond lengths over small molecules and two cyclophanes in the second to last column, and it presents the average of the MUEs for MGBL24 and CID15 in the last column. Table 14 shows that the best performers are M06-L and PBEh, followed by M08-HX, M08-SO, SOGGA, M05-2X, and M06-2X.

***Table 14.*** Statistical Errors (Å) for Bond Lengths

| method | rung | $X$ | $\mu$ | MGBL24[a] | CID15[b] | MGBL34[c] | AMUE[d] |
|---|---|---|---|---|---|---|---|
| M06-L | 3 | 0 | 0.2678 | 0.007 | 0.008 | 0.006 | 0.007 |
| PBEh | 4 (HG) | 25 | 0.1955 | 0.009 | 0.007 | 0.008 | 0.008 |
| M08-HX | 4 (HM) | 52.23 | 0.2696 | 0.013 | 0.007 | 0.011 | 0.010 |
| M08-SO | 4 (HM) | 56.79 | 0.1235 | 0.014 | 0.006 | 0.011 | 0.010 |
| SOGGA | 2 | 0 | 0.1235 | 0.014 | 0.007 | 0.011 | 0.010 |
| M05-2X | 4 (HM) | 56 | 0.1889 | 0.016 | 0.006 | 0.013 | 0.011 |
| M06-2X | 4 (HM) | 54 | 0.1881 | 0.015 | 0.006 | 0.012 | 0.011 |
| TPSSh | 4 (HM) | 10 | 0.1235 | 0.012 | 0.012 | 0.011 | 0.012 |
| TPSS | 3 | 0 | 0.1235 | 0.015 | 0.018 | 0.015 | 0.017 |
| B97−1 | 4 (HG) | 21 | 0.1654 | 0.010 | 0.023 | 0.010 | 0.017 |
| PBE | 2 | 0 | 0.2195 | 0.014 | 0.020 | 0.013 | 0.017 |
| BMK | 4 (HM) | 42 | 1.1112 | 0.016 | 0.020 | 0.015 | 0.018 |
| B3LYP | 4 (HG) | 20 | 0.2222 | 0.011 | 0.026 | 0.011 | 0.018 |

[a] MUE of the MGBL24 database of Table S4. [b] MUE of the 15 bond lengths in Tables S5 and S6. [c] Average of MGBL24 and CID15. [d] MUE of the 34 bond distances in MGBL24 and CID15, that is all 24 distances in MGBL24 and the 10 smallest distances of CID15 (the other five distances in CID15 are nonbonded distances).

***Table 15.*** Scale Factor and Statistical Errors ($cm^{-1}$) for the F38 Database

| | B3LYP | M08-SO | M08-HX | M06-2X | M05-2X |
|---|---|---|---|---|---|
| MSE | 8 | 20 | 44 | 49 | 64 |
| MUE | 31 | 52 | 56 | 56 | 70 |
| scale factor | 0.998 | 0.995 | 0.984 | 0.982 | 0.975 |
| MSE after scaling | 4 | 9 | 9 | 9 | 8 |
| MUE after scaling | 31 | 51 | 49 | 45 | 44 |

*5.2.8. Frequencies.* M06-2X and the two M08 functionals do not improve on B3LYP for frequencies in small molecules, as shown in Tables S7 and 15. We also optimized a scale factor for harmonic frequencies for each of the tested functionals, with the optimization being to improve the harmonic frequencies, as in ref 28. After scaling, the MUE

***Table 16.*** Vertical Excitation Energies (VEE) for $H_2O$ (kcal/mol)[a]

| method | $^1A_1 \rightarrow {}^3B_1$ |
|---|---|
| CCSD(T)/aug-cc-pVTZ | 167.58 |
| CCSD(T)/aug-cc-pVQZ | 168.68 |
| CCSDT/aug-cc-pVTZ | 167.37 |
| $CCSDT(2)_Q$/aug-cc-pVTZ | 167.44 |
| Q(TZ)[b] | −0.13 |
| CCSD(T)/aug-cc-pVQZ+Q(TZ) | 168.55 |

[a] At the experimental[133] geometry: $r_{OH} = 0.957$, $\theta_{HOH} = 104.5$ [b] Q(TZ) = VEE($CCSDT(2)_Q$/aug-cc-pVTZ) − VEE(CCSD(T)/aug-cc-pVTZ).

of M06-2X and M08-HX decrease by 11 and 7 $cm^{-1}$, respectively.

*5.2.9. Multiplicity-Changing Excitation Energies.* For transitions to electronic states with a different multiplicity from the ground state, we calculated the excitation energy by performing self-consistent-field (SCF) calculations on both states. The reference vertical excitation energy for $H_2O$ is calculated by using the CCSD(T)/aug-cc-pVQZ+Q(TZ) method, where Q(TZ) stands for the quadruple excitation correlation contributions calculated at the $CCSDT(2)_Q$ level[181] of theory. Table 16 summarizes the WFT results for the vertical excitation of $H_2O$, with some results from the literature.

Besides the best estimates and the M06-2X, M08-HX, M08-SO, B3LYP, and M05-2X results, we also present the multiplicity-changing excitation energies for nine other methods in Table 17. Two of the methods, BMK and B97-3, are chosen for comparison because in previous tests[28] on 41 diverse excitations (23 of which conserve multiplicity and 18 of which do not) these two methods showed the best mean performance of any of the non-Minnesota functionals that were tested. The other methods in Table 17 illustrate the effect of introducing empirical parameters for exchange and correlation in the set of functionals built on B88 exchange, HF exchange, and LYP correlation.

Table 17 shows that the new M08-HX functional is the best performer for the multiplicity-changing excitation energy excitation energy database, followed by B97-3, BMK, M08-SO, and M06-2X. The B3LYP*, B3LYP, BLYP, HFLYP, B88, and HF functionals in Table 17 show the effects of several different ways to mix and scale the components of a hybrid GGA. HF has 100% HF exchange with no correlation energy; B88 has 100% B88 exchange with no correlation energy. HFLYP has 100% HF exchange with 100% LYP correlation energy; BLYP has 100% B88 exchange with 100% LYP correlation energy. B3LYP has 20% HF exchange, 72% B88 exchange, and 80% LSDA exchange with 81% LYP correlation energy, and B3LYP* has 15% HF exchange, 72% B88 exchange, and 85% LSDA exchange with 81% LYP correlation energy. The results obtained for these six functionals show several features that merit further consideration. First of all, HF exchange is often said to relatively overstablize high-spin states, and Table 17 shows that, in comparison to experiment, HFLYP underestimates the excitation energy to a high-multiplicity state in eight out of nine cases. Furthermore, as compared to B88, HF underestimates the excitation energy to a high-spin state in seven of nine cases. These results confirm expectations. Next, however, compare B88 to HFLYP. Although exchange has usually been considered to be the key to correct multiplicity ordering and spin-state splitting,[174,182−191] the difference in MUE between HFLYP and HF is larger than the difference between B88 and HF, showing the importance of differential dynamical correlation for spin-state splitting. Another "surprise" is found by comparing B3LYP to B3LYP*, a functional that was developed by Reiher and co-workers[174,182] especially to improve multiplicity-changing excitation energies by decreasing the weighting of HF exchange since the HF theory underestimates spin-state splittings. Table 17 shows that, as compared to B3LYP, sometimes B3LYP* predicts higher spin-state splitting and sometimes lower; on average it is only slightly more accurate than B3LYP.

*5.2.10. Discussion of $\mu$.* In general the orbitals, like the Hartree−Fock exchange functional, are nonlocal functions of the density, so a local functional of the orbitals brings in some nonlocal information and is sometimes called semilocal.[192] However in our classification (which is also used by some other workers—there is no consensus on the language), any functional that depends only on local values of the spin densities, their gradient magnitudes, and the spin kinetic energy densities is called local. Functionals that are not local are called nonlocal or hybrid. To the best of our knowledge, $\mu$ values have not been presented previously for hybrid functionals, so the comparison of $\mu$ values in Table 3 merits some discussion.

It has been known for a long time that the value of $\mu$ can be helpful in understanding the performance of GGAs, and it has sometimes been stated that functionals with the gradient-expansion value (that is $\mu_{GE} = 10/81 \approx 0.1246$) of $\mu$ should be more accurate for solids and surfaces, whereas those with values about twice as high should be more accurate for free atoms and small molecules.[192−196] Our recent study[8] provided a more nuanced conclusion, namely that using $\mu \approx \mu_{GE}$ leads to better accuracy for interatomic spacings not only in solids but also in molecules, at least for bond lengths that do not involve hydrogens, whereas $\mu \approx 2\ \mu_{GE}$ leads to better accuracy not only for atomization energies of molecules but also for barrier heights of chemical reactions and cohesive energies of solids.

It is impossible for a single GGA to be highly accurate for both interatomic spacings and energetics. Adding orbital dependencies such as Hartree−Fock exchange and kinetic energy densities can ameliorate this situation. Comparing eq 3 to 14 shows that $\tau$ and $s$ bring in similar information in the slow-varying density limit, but the use of $\tau$ at finite values of $s$ allow one to distinguish different kinds of electron density regions that have the same $s$ and $\rho$. Because the explicit dependence on spin kinetic energy densities allows one to distinguish different regions with the same $s$ and $\rho$, the performance of the hybrid meta functionals does not correlate with $\mu$ in the same way as for GGAs. One might ultimately prefer a functional with $\mu = \mu^{GE}$ (as in M08-SO, TPSS, TPSSh, and SOGGA), but so far only M08-SO performs as well for chemistry as do M05-2X, M06-2X, and

***Table 17.*** Multiplicity-Changing Excitation Energies (kcal/mol)[a]

| transition | Mg $^1S \rightarrow {}^3P$ | Be $^1S \rightarrow {}^3P$ | $H_2CO$ $^1A_1 \rightarrow {}^3A_2$ | BeH $^2\Sigma_1 \rightarrow {}^4\Pi_2$ | CO $^1\Sigma^+ \rightarrow {}^3\Pi$ | $H_2O$ $^1A_1 \rightarrow {}^3B_1$ | $N_2$ $^1\Sigma_g^+ \rightarrow {}^3\Sigma_u^+$ | $NO_2$[b] $^2A_1 \rightarrow {}^4A_2$ | SiO[b] $^1\Sigma^+ \rightarrow {}^3\Sigma^+$ | MSE | MUE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| best estimate | 62.47 | 62.84 | 80.71 | 134.67 | 145.74 | 168.55 | 178.72 | 83.30 | 96.15 | | |
| M08-HX | 64.66 | 60.63 | 80.32 | 139.85 | 145.99 | 170.29 | 186.91 | 81.50 | 95.78 | 1.42 | 2.48 |
| B97−3 | 64.09 | 56.13 | 77.83 | 136.94 | 142.05 | 168.92 | 180.76 | 79.26 | 93.25 | −1.55 | 2.95 |
| BMK | 61.11 | 52.97 | 79.19 | 134.13 | 145.84 | 170.08 | 185.41 | 77.99 | 96.94 | −1.06 | 3.08 |
| M08-SO | 65.32 | 61.90 | 80.81 | 143.00 | 146.60 | 168.13 | 187.26 | 79.36 | 98.05 | 1.92 | 3.10 |
| M06−2X | 69.86 | 63.12 | 80.82 | 143.14 | 142.39 | 170.18 | 185.77 | 79.18 | 95.73 | 1.89 | 3.65 |
| B3LYP* | 64.35 | 56.55 | 76.13 | 139.46 | 140.91 | 164.04 | 176.67 | 79.88 | 92.46 | −2.52 | 4.00 |
| B3LYP | 64.11 | 56.61 | 75.86 | 140.01 | 140.87 | 163.43 | 177.35 | 77.81 | 92.41 | −2.74 | 4.29 |
| BLYP | 65.46 | 56.81 | 76.53 | 140.28 | 140.82 | 163.58 | 173.91 | 81.66 | 91.24 | −2.54 | 4.45 |
| M06-L | 60.76 | 53.39 | 77.09 | 81.98 | 91.64 | 139.18 | 136.33 | 170.56 | 174.50 | −3.08 | 4.53 |
| M05-2X | 75.03 | 65.98 | 79.69 | 141.16 | 147.30 | 174.34 | 186.74 | 79.18 | 98.72 | 3.89 | 5.03 |
| M06-HF | 70.59 | 70.12 | 85.65 | 147.79 | 147.41 | 170.50 | 191.95 | 77.08 | 98.31 | 5.14 | 6.52 |
| HFLYP | 60.23 | 56.57 | 65.56 | 141.12 | 138.23 | 155.94 | 158.77 | 36.35 | 93.59 | −11.86 | 13.30 |
| B88 | 46.47 | 37.62 | 70.80 | 113.97 | 127.42 | 146.60 | 167.68 | 70.52 | 79.16 | −16.99 | 16.99 |
| HF | 41.20 | 37.28 | 58.08 | 114.87 | 121.83 | 137.67 | 148.42 | 35.46 | 80.02 | −26.48 | 26.48 |

[a] The reference vertical excitation energy for $H_2O$ is the result from CCSD(T)/aug-cc-pVQZ + Q(TZ) in Table 21. See Section 2.2.12 for the source of reference data for other molecules or atoms. [b] These are adiabatic excitation energies; others are vertical excitation energies.

M08-HX; these three functionals have $\mu$ values that are 1.52−2.18 times larger than $\mu_{GE}$.

Table 3 shows that M08-HX has a larger value of $\mu$ than M05-2X or M06-2X (0.2696 vs 0.1885−0.1889), whereas previous work[8] shows that for GGAs, functionals with even smaller $\mu$ (0.1235) predict more accurate nonhydrogenic bond distances. Nevertheless our strategy of allowing a more flexible functional form to allow the resulting functional to be simultaneously more accurate for both energetics and bond distances did succeed in that M08-HX, although fit only to energetics and although more accurate on average than either M05-2X or M06-2X for energetics, is also significantly more accurate for typical bond distances (see Table 14). In fact, functionals with high HF exchange are usually expected to be less accurate than low-$X$ functionals for bond distances, but M08-HX outperforms not only TPSS and PBE, which have $X = 0$, but also SOGGA, which not only has $X = 0$ but also has $\mu = \mu^{GE}$ which is known from previous work to be associated with good accuracy for lattice constants and nonhydrogenic bond distances.

## 6. Concluding Remarks

This paper presents two new hybrid meta-GGA exchange-correlation functionals, M08-2X and M08-SO, for main-group thermochemistry, thermochemical kinetics, and noncovalent interactions. The new M08-HX functional has an improved functional form as compared to our previous M06-2X and M05-2X functional forms. The M08 functional form rigorously enforces the UEG limit and avoids the use of a self-correlation correction term, which sometimes causes difficulties in the SCF iterations. The M08-HX, M08-SO, M06-2X, and M05-2X functionals have been comparatively assessed against 164 energetic test data, 39 bond lengths, and 38 frequencies outside of the training set.

Before summarizing what we learn from the present research, we remind the readers that since M08-2X, M06-2X, and M05-2X have high Hartree−Fock exchange, they are not parametrized to be suitable for studying many problems in transition metal chemistry or other problems with high multireference character.

From the assessment, we draw the following conclusions:

1) The limit of accuracy of a global hybrid meta-GGA for our training set of 267 molecular data is about 0.75 kcal/mol.

2) M08-HX, M08-SO, and M06-2X perform very well for a combination of main-group thermochemistry, kinetics, and noncovalent interactions.

3) M08-HX, M08-SO, and M06-2X give good performances for the noncovalent interactions in large uracil trimers and for conformational energies in small aromatic peptides.

4) M08-HX, M08-SO, M06-2X, and M05-2X do not improve upon B3LYP for bond lengths in small molecules, but they perform well in predicting the bond lengths in cyclophanes, for which B3LYP fails.

5) M08-2X, M08-SO, M06-2X, and M05-2X do not improve upon B3LYP for frequencies.

6) M08-SO is considerably more accurate for main-group thermochemistry than any previously available functional with the correct second-order behavior in the regime of slowly varying density.

7) The new M08-HX functional has the best performance of tested functionals for several of the databases, in particular, main-group atomization energies, large-molecule atomization energies, electron affinities, hydrogen-transfer barrier heights, heavy-atom transfer barrier heights, new benchmark barrier heights, noncovalent interaction energies in uracil trimers, and multiplicity-changing excitation energies.

**Acknowledgment.** This work was supported in part by the Air Force Office of Scientific Research (orbital-dependent density functionals), by the National Science Foundation under grant no. CHE07-04974 (complex systems), by the Office of Naval Research under award number N00014-05-0538 (software tools), and by a Molecular Science Computing Facility Computational Grand Challenge grant at the Environmental Molecular Science Laboratory of Pacific Northwestern National Laboratory.

**Supporting Information Available:** Isomerization energies (Table S1), results for a double-$\zeta$ basis set (Tables

S2 and S3), bond lengths (Table S4), internuclear distances (Tables S5 and S6), frequencies (Table S7), and Cartesian coordinates (Tables S8-S11). This material is available free of charge via the Internet at http://pubs.acs.org.

## References


(1) Scuseria, G. E.; Staroverov, V. N. In *Theory and Application of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 669.

(2) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, 1133.

(3) Langreth, D. C.; Mehl, M. J. *Phys. Rev. B* **1983**, *28*, 1809.

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

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

(6) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J. *Phys. Rev. B* **1992**, *46*, 6671.

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

(8) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2008**, *128*, 184109.

(9) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

(10) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

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

(12) Gruning, M.; Gritsenko, O.; Baerends, E. J. *J. Phys. Chem. A* **2004**, *108*, 4459.

(13) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

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

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

(16) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439.

(17) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(18) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(19) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

(20) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.

(21) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.

(22) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

(23) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

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

(25) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.

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

(27) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

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

(29) Becke, A. D. *Int. J. Quantum Chem.* **1983**, *23*, 1915.

(30) Becke, A. D. *J. Chem. Phys.* **1998**, *109*, 2092.

(31) Becke, A. D. *J. Chem. Phys.* **2000**, *112*, 4020.

(32) Schmider, H. L.; Becke, A. D. *THEOCHEM* **2000**, *527*, 51.

(33) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(34) Woodcock, H. L.; Schaefer, H. F.; Schreiner, P. R. *J. Phys. Chem. A* **2002**, *106*, 11923.

(35) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.

(36) Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624.

(37) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.

(38) Carlier, P. R.; Deora, N.; Crawford, T. D. *J. Org. Chem.* **2006**, *71*, 1592.

(39) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.

(40) Schreiner, P. R.; Fokin, A. A., Jr.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.

(41) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.

(42) Izgorodina, E. I.; Coote, M. L. *J. Phys. Chem A.* **2006**, *110*, 2486.

(43) Izgorodina, E. I.; Coote, M. L. *Chem. Phys.* **2006**, *324*, 96.

(44) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42.

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

(46) Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217.

(47) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 10754.

(48) Paier, J.; Marsman, M.; Kresse, G. *J. Chem. Phys.* **2007**, *127*, 24103.

(49) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.

(50) Paesani, F.; Gianturco, F. A.; Lewerenz, M. *J. Chem. Phys.* **1999**, *111*, 6897.

(51) Ortmann, F.; Schmidt, W. G.; Bechstedt, F. *Phys. Rev. Lett.* **2005**, *95*, 186101.

(52) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(53) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(54) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 174104.

(55) Puzder, A.; Dion, M.; Langreth, D. C. *J. Chem. Phys.* **2006**, *124*, 164105.

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

(57) Morgado, C. A.; McNamara, J. P.; Hillier, I. H.; Burton, N. A.; Vincent, M. A. *J. Chem. Theory Comput.* **2007**, *3*, 1656.

(58) Tapavicza, E.; Lin, I.-C.; Lilienfeld, O. A. v.; Tavernelli, I.; Coutinho-Neto, M. D.; Rothlisberger, U. *J. Chem. Theory Comput.* **2007**, *3*, 1673.

(59) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. *J. Phys. Chem. A* **2007**, *111*, 1146.

(60) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.

(61) Grimme, S.; Antony, J.; Schwabe, T.; Mueck-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, *5*, 741.

(62) Morgado, C.; Vincent, M. A.; Hillier, I. H.; Shan, X. *Phys. Chem. Chem. Phys.* **2007**, *9*, 448.

(63) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

(64) Grimme, S.; Mueck-Lichtenfeld, C.; Antony, J. *J. Phys. Chem. C* **2007**, *111*, 11199.

(65) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.

(66) Savin, A. In *Recent Developments and Applications of Modern Density Functional Theory*; Seminario, J., Ed.; Elsevier: Amsterdam, 1996; p 327.

(67) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2003**, *118*, 8207.

(68) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.

(69) Gerber, I. C.; Ángyán, J. G. *Chem. Phys. Lett.* **2005**, *415*, 100.

(70) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109. /1.

(71) Song, J.-W.; Tokura, S.; Sato, T.; Watson, M. A.; Hirao, K. *J. Chem. Phys.* **2007**, *127*, 154109.

(72) Goll, E.; Werner, H.-J.; Stoll, H. *Chem. Phys.* **2008**, *346*, 257.

(73) Chai, J.-D.; Martin, H.-G. *J. Chem. Phys.* **2008**, *128*, 84106.

(74) Jaramillo, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 1068.

(75) Arbuznikov, A. V.; Kaupp, M.; Bahmann, H. *J. Chem. Phys.* **2006**, *124*, 204102.

(76) Bahmann, H.; Rodenberg, A.; Arbuznikov, A. V.; Kaupp, M. *J. Chem. Phys.* **2007**, *126*, 11103.

(77) Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2007**, *127*, 164117.

(78) Kaupp, M.; Bahmann, H.; Arbuznikov, A. V. *J. Chem. Phys.* **2007**, *127*, 194102.

(79) Arbuznikov, A. V.; Kaupp, M. *Chem. Phys. Lett.* **2007**, *440*, 160.

(80) Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *128*, 84111.

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

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

(83) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 34108.

(84) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398.

(85) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D. *J. Phys. Chem A* **2008**, *112*, 3.

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

(87) Perdew, J. P.; Schmidt, K. In *Density Functional Theory and Its Applications to Materials*; Van-Doren, V., Alsenoy, C. V., Geerlings, P., Eds.; American Institute of Physics: New York, 2001; p 1.

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

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

(90) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.

(91) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.

(92) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.

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

(94) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2006**, *8*, 5753.

(95) Zhao, Y.; Truhlar, D. G. *J. Org. Chem.* **2007**, *72*, 295.

(96) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289.

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

(98) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 1095.

(99) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 6860.

(100) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 3976.

(101) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R. *Org. Lett.* **2007**, *9*, 1851.

(102) Rokob, T. A.; Hamza, A.; Papai, I. *Org. Lett.* **2007**, *9*, 4279.

(103) Shields, A. E.; van Mourik, T. *J. Phys. Chem A* **2007**, *111*, 13272.

(104) Pietra, F. *J. Phys. Org. Chem.* **2007**, *20*, 1102.

(105) Santoro, F.; Barone, V.; Improta, R. *J. Comput. Chem.* **2008**, *29*, 957.

(106) Brancato, G.; Rega, N.; Barone, V. *Phys. Rev. Lett.* **2008**, *100*, 107401.

(107) Improta, R.; Berisio, R.; Vitagliano, L. *Protein Sci.* **2008**, *17*, 955.

(108) Ellingson, B. A.; Truhlar, D. G. *J. Am. Chem. Soc.* **2007**, *129*, 12765.

(109) Hayama, T.; Baldridge, K. K.; Wu, Y.-T.; Linden, A.; Siegel, J. S. *J. Am. Chem. Soc.* **2008**, *130*, 1583.

(110) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 4061.

(111) Zhao, Y.; Truhlar, D. G. *J. Am. Chem. Soc.* **2007**, *129*, 8440.

(112) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2813.

(113) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

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

(115) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012 (E) **2006**, *110*, 4942.

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

(117) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. F. *Phys. Rev. A* **1993**, *47*, 3649.

(118) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.

(119) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem A* **2006**, *110*, 13551.

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

(121) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *10*, 2497.

(122) Kabelác, M.; Sherer, E. C.; Cramer, C. J.; Hobza, P. *Chem. Eur. J.* **2006**.

(123) Pascal, R. A.; Winans, C. G.; Engen, D. V. *J. Am. Chem. Soc.* **1989**, *111*, 3007.

(124) Grimme, S. *Chem. Eur. J.* **2004**, *10*, 3423.

(125) Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *4*, 297.

(126) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.

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

(128) Martin, J. M. L.; Oliveira, G. D. *J. Chem. Phys.* **1999**, *111*, 1843.

(129) Zheng, J. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem A* **2007**, *111*, 4632.

(130) Gonzales, J. M.; Allen, W. D.; Schaefer, H. F., III *J. Phys. Chem. A* **2005**, *109*, 10613.

(131) Troya, D. *J. Phys. Chem A* **2007**, *111*, 10745.

(132) *NIST Computational Chemistry Comparison and Benchmark Database*; Johnson, R. D., III, Ed.; 2006. http://srdata.nist.gov/cccbdb/ (accessed July 2008).

(133) Handy, N. C.; Tozer, D. J. *Mol. Phys.* **1998**, *94*, 707.

(134) http://webbook.nist.gov/chemistry/ (accessed July 2008).

(135) Pitarch-Ruiz, J.; Sánchez-Marín, J.; Velasco, A. M. *J. Comput. Chem.* **2008**, *29*, 523.

(136) Ben-Shlomo, S. B.; Kaldor, U. *J. Chem. Phys.* **1990**, *92*, 3680.

(137) Nielsen, E. S.; Jorgensen, P.; Oddershede, J. *J. Chem. Phys.* **1980**, *73*, 6238.

(138) Clouthier, D. J.; Ramsay, D. A. *Annu. Rev. Phys. Chem.* **1983**, *34*, 31.

(139) Bera, P. P.; Yamaguchi, Y.; Henry, F.; Schaefer, I. *J. Chem. Phys.* **2007**, *127*, 174303.

(140) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

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

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

(143) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

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

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

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

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

(148) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 84108.

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

(150) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418. 1987, 86, 3760 (E).

(151) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.

(152) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03; Revision E.01 ed.*; Gaussian, Inc.: Pittsburgh, PA, 2008.

(153) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module - Version 4.0beta*; University of Minnesota: Minneapolis, MN, 2006.

(154) Bylaska, E. J.; Jong, W. A. d.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Aprà, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; Lenthe, J. v.; Wong, A.; Zhang, Z. *NWChem A Computational Chemistry Package for Parallel Computers; Version 5.1 ed.*; Pacific Northwest National Laboratory: Richland, WA, U.S.A., 2007.

(155) Oliver, G. L.; Perdew, J. P. *Phys. Rev. A* **1979**, *20*, 397.

(156) Antoniewicz, P. R.; Kleinman, L. *Phys. Rev. B* **1985**, *31*, 6779.

(157) Hammer, B.; Hansen, L. B.; Norskov, J. K. *Phys. Rev. B* **1999**, *59*, 7413.

(158) Thomas, L. H. *Proc. Camb. Philos. Soc.* **1927**, *23*, 542.

(159) Fermi, E. *Z. Phys.* **1928**, *48*, 73.

(160) Brack, M.; Jennings, B. K.; Chu, Y. H. *Phys. Lett. B* **1976**, *65*, 1.

(161) Stoll, H.; Pavkidou, C. M. E.; Preuss, H. *Theor. Chim. Acta* **1978**, *49*, 143.

(162) Gori-Giorgi, P.; Perdew, J. P. *Phys. Rev. B* **2004**, *69*, 041103.

(163) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 201102.

(164) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2007**, *126*, 104102.

(165) Gräfenstein, J.; Izotov, D.; Cremer, D. *J. Chem. Phys.* **2007**, *127*, 214103.

(166) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(167) Svendesen, P. S.; von Barth, U. *Phys. Rev. B* **1996**, *54*, 17402.

(168) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *Phys. Rev. B* **2004**, *69*, 75102.

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

(170) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(171) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

(172) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(173) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.

(174) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48.

(175) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.

(176) Valentin, C. D.; Pacchioni, G.; Bredow, T.; Dominguez-Ariza, D.; Illas, F. *J. Chem. Phys.* **2002**, *117*, 2299.

(177) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(178) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2008**, *128*, 124105.

(179) Pasacal, R. A. *J. Phys. Chem. A* **2001**, *105*, 9040.

(180) Pascal, R. A. *Eur. J. Org. Chem.* **2004**, *18*, 3763.

(181) Hirata, S.; Fan, P.-D.; Auer, A. A.; Nooijen, M.; Piecuch, P. *J. Chem. Phys.* **2004**, *121*, 12197.

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

(183) Poli, R.; Harvey, J. N. *Chem. Soc. Rev.* **2003**, *32*, 1.

(184) Paulsen, H.; Trautwein, A. X. *J. Phys. Chem. Solids* **2004**, *65*, 793.

(185) Fouqueau, A.; Mer, S.; Casida, M. E.; Daku, L.; Max, L.; Hauser, A.; Mineva, T.; Neese, F. *J. Chem. Phys.* **2004**, *120*, 9473.

(186) Paulsen, H.; Trautwein, A. X. *Top. Curr. Chem.* **2004**, *235*, 197.

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

(188) Harvey, J. N. *Struct. Bonding (Berlin)* **2004**, *112*, 151.

(189) Liao, M.-S.; Watts, J. D.; Huang, M.-J. *J. Comput. Chem.* **2006**, *27*, 1577.

(190) Rong, C.; Lian, S.; Yin, D.; Shen, B.; Zhong, A.; Bartolotti, L.; Liu, S. *J. Chem. Phys.* **2006**, *125*, 174102.

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

(192) Kurth, S.; Perdew, J. P.; Blaha, P. *Int. J. Quantum Chem.* **1999**, *75*, 889.

(193) Perdew, J. P.; Constantin, L. A.; Sagvolden, E.; Burke, K. *Phys. Rev. Lett.* **2006**, *97*, 223002.

(194) Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Ruzsinszky, A.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 244107.

(195) Madsen, G. K. H. *Phys. Rev. B* **2007**, *75*, 195108.

(196) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 136406.

CT800246V

JCTC Journal of Chemical Theory and Computation

# Structures, Bonding, and One-Bond B−N and B−H Spin−Spin Coupling Constants for a Series of Neutral and Anionic Five-Membered Rings Containing BN Bonds

Manuel Yáñez,*,† Otilia Mó,† Ibon Alkorta,‡ and Janet E. Del Bene§

*Departamento de Química, C-9, Universidad Autónoma de Madrid, Cantoblanco, 28049-Madrid, Spain, Instituto de Química Médica (CSIC), Juan de la Cierva, 3, E-28006-Madrid, Spain, and Department of Chemistry, Youngstown State University, Youngstown, Ohio 44555*

Received July 11, 2008

**Abstract:** The structures and bonding of a series of five-membered rings with BN bonds $C_xN_yB_zH_5$ (x + y + z = 5) and their most stable deprotonated anions $C_xN_yB_zH_4^-$ as well as anionic rings $C_xN_yB_zH_5^-$ have been investigated at the MP2/6−311++G(d,p) level of theory. The great majority of these rings present BN bond orders close to that found in borazine, suggesting that there is substantial electron delocalization in these rings. This observation is also supported by both NBO and ELF analyses. Ab initio equation-of-motion coupled-cluster singles and doubles (EOM-CCSD) calculations have also been performed to obtain the $^{15}N-^{11}B$ and $^{1}H-^{11}B$ spin−spin coupling constants. For neutral systems, the former range from −10 to −35 Hz, thereby bracketing the value of $^{1}J(B-N)$ for borazine, which is −29 Hz. $^{1}J(B-N)$ spans an even greater range in the anions, from −3 to −36 Hz. The absolute value of $^{1}J(B-N)$ decreases upon deprotonation if coupling involves the deprotonated nitrogen or a boron atom bonded to the deprotonated N. $^{1}J(B-H)$ always decreases upon nitrogen deprotonation.

## Introduction

Boron is an element with quite unusual bonding properties[1,2] which arise from its intrinsic electron-deficient character and are reflected in the versatility of its chemistry. Many boron derivatives are widely used in synthesis because they are very good electrophiles. However, a novel synthesis recently produced a diazaborole five-membered ring with BN bonds and B acting as a nucleophilic center.[3] Other five-membered rings containing BN bonds have also been synthesized and characterized as suitable ligands.[4,5] They can replace the usual cyclopentadienyl ligand and form face-on sandwich complexes.[5] However, while the all-carbon ligands display almost exclusively $\eta^5$ coordination, the BN-containing analogues exhibit variability in their coordination manner, and $\eta^1$, $\eta^2$, $\eta^3$, and $\eta^4$ coordination modes have been assigned.[5] The variability of the coordination modes appears to be closely related to BN bonding characteristics and electron delocalization due to the presence of BN bonds in the rings. In rings containing a NN bond instead, electron delocalization seems to be restricted to the NBCBN skeleton with no participation of the NN linkage.

We have recently been involved in a systematic study of the BN bond in different environments,[2,6−8] paying particular attention to bonding characteristics and $^{11}B-^{15}N$ coupling constants. In the present paper we present the results of the next logical extension of our studies to five-membered BN-containing rings. For this study we have considered both neutral and anionic rings containing at least one B−N bond and fully saturated valency. Only the most stable anion derived from one of the neutral rings by deprotonation has been included.

## Computational Details

The geometries of the BN rings were optimized at the MP2/6−311++G(d,p) level. Harmonic vibrational frequencies

* Corresponding author e-mail: manuel.yanez@uam.es.
† Universidad Autónoma de Madrid.
‡ Instituto de Química Médica (CSIC).
§ Youngstown State University.

10.1021/ct8002699 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/24/2008

were calculated at the same level to confirm that the stationary points found are local minima on the corresponding potential energy surfaces. The bonding in these rings has been analyzed by means of the NBO partitioning technique,[9] the electron localization function (ELF) theory,[10] and the atoms in molecules (AIM) theory.[11] The NBO approach examines the bonding in terms of localized hybrids and lone pairs obtained as local block eigenvectors of the one-particle density matrix as well as in terms of the interactions between occupied and unoccupied MOs through a second-order perturbation analysis of the Fock matrix. These NBO calculations were carried out allowing for the detection of three-center bonding. The Wiberg bond orders were also evaluated in the framework of this approach. By means of the AIM theory we have located for each ring the different bond critical points (BCPs) that together with the bond paths define the molecular graph. At each BCP the electron density was also evaluated since it usually provides information about the nature and relative strengths of the bonds with which the BCP is associated. ELF[10] is a function which becomes large in regions of space where electron pairs, either bonding or lone pairs, are localized. Usually an ELF value around 0.80 defines isosurfaces which provide clear pictures of the regions of electron localization or attraction basins which may be related to key bonding concepts, such as core, valence, and lone-pair regions, while their populations have been related to bond order. ELF grids and basin integrations have been evaluated with the TopMod package.[12]

The one-bond $^{11}B-^{15}N$ and $^{11}B-^{1}H$ coupling constants were computed using the equation-of-motion coupled-cluster singles and doubles (EOM-CCSD)[13−16] method in the CI(configuration interaction)-like approximation with all electrons correlated. These calculations were carried out using the Ahlrichs[17] qzp basis set on C and N, the qz2p basis on H atoms bonded to B, and the cc-pVDZ basis set on the remaining H atoms.[18,19] Since an Ahlrichs qzp basis is not available for B, a new "hybrid" basis was constructed for this atom from the boron cc-pV5Z basis for s orbitals, the boron cc-pVQZ for p orbitals, and a single set of d polarization functions and used in previous studies[6−8] of systems containing BN bonds. The total coupling constant J is a sum of four terms: paramagnetic spin−orbit (PSO), diamagnetic spin−orbit (DSO), Fermi-contact (FC), and spin-dipole (SD). For selected molecules and anions, all terms were evaluated. However, because of the size and low-computational symmetry of the BN rings, only the FC term was evaluated and used to approximate total J for the majority of molecules and ions. That the FC term is a good approximation to total J for B−N and B−H coupling constants was observed previously in the study of the six-membered borazine rings[6] and is also evident from the data of Table 1. In general, the FC term overestimates the absolute value of $^{1}J(B-N)$ by about 2 Hz, the approximate contribution from the PSO term, and is an excellent approximation to $^{1}J(B-H)$. Hence, in this paper, the FC term will be used to approximate total J. The level of theory used for these calculations has been shown to reproduce the experimental one-bond C−N coupling constants in the prototypical six-membered rings pyridine and pyridinium and the C−P coupling constant in phosphinine.[20] In addition, the computed B−H coupling constant in newly synthesized diazaborole is in agreement with the experimental value of $^{1}J(B-H)$.[8] Finally, good agreement between experimental and computed coupling constants has been found for a series of molecules $H_mX\text{-}YH_n$ and selected F-derivatives.[21] Coupling constant calculations were carried out using ACES II[22] on the Itanium cluster at the Ohio Supercomputer Center.

***Table 1.*** B−N and B−H Total Coupling Constants (J, Hz) and Components of J for Selected Molecules and Ions

| species | pair | PSO | DSO | FC | SD | J |
|---|---|---|---|---|---|---|
| **C2N2B_1** | N1−B2 | 2.34 | −0.11 | −32.40 | −0.27 | −30.45 |
| **C2NB2_1** | N1−B2 | 2.24 | −0.08 | −25.75 | −0.10 | −23.70 |
| **C2NB2_1_A** | N1−B2 | 2.80 | −0.07 | −15.64 | −0.09 | −13.00 |
| **N3B2_2_A** | N1−B2 | 2.43 | −0.10 | −19.19 | −0.37 | −17.24 |
| **N2B3_1** | N1−B2 | 1.78 | −0.09 | −29.62 | −0.10 | −28.03 |
| | N3−B4 | 2.82 | −0.08 | −20.49 | −0.12 | −17.87 |
| **C2N2B_1** | B2−H7 | −0.82 | 0.99 | 145.30 | 0.31 | 145.77 |
| **C2NB2_1** | B2−H7 | −1.22 | 0.74 | 128.89 | 0.39 | 128.80 |
| **C2NB2_1_A** | B2−H6 | −1.15 | 0.79 | 108.88 | 0.20 | 108.71 |
| **N3B2_2_A** | B2−H7 | −0.40 | 0.78 | 104.95 | 0.11 | 105.44 |
| **N2B3_1** | B2−H7 | −1.12 | 0.89 | 134.37 | 0.40 | 134.55 |
| | B4−H9 | −1.31 | 0.64 | 122.68 | 0.28 | 122.28 |

## Structures and Bonding

A total of 32 neutral and anionic five-membered BN-containing rings with the general formulas $C_xN_yB_zH_5$ ($x + y + z = 5$), their most stable deprotonated anions $C_xN_yB_zH_4^-$, and anionic rings $C_xN_yB_zH_5^-$ have been investigated, although calculations on additional anions obtained by deprotonation of the neutral rings were carried out so that the most stable anion could be identified. A systematic identification scheme has been developed so that the neutral molecules are listed in order of decreasing number of carbon atoms, and for a given number of carbons, decreasing number of nitrogen atoms **CxNyBz**. If a particular compound has more than one isomer, these are listed in order of relative stability, indicated by a number. For example, **N3B2_1** is more stable than **N3B2_2**. Anionic species have the letter **A** added to the label, as in **C2N2B_1_A**. Figure 1 presents the five-membered rings investigated in this study and illustrates the atom numbering adopted so that each ring has an N1−B2 bond. Usually, N1 is also the deprotonated nitrogen. An anion **CN3B_A2** originally considered for inclusion was eliminated from this study because it presents a RHF-UHF instability, thereby invalidating both MP2 and EOM-CCSD calculations.

The Wiberg bond orders for HBNH (2.140), $H_2BNH_2$ (1.257), $H_3BNH_3$ (0.625), and borazine (1.011) as paradigmatic examples of triple, double, single, and aromatic BN bonds, respectively, are suitable references for the discussion which follows. The Wiberg bond orders (BOs) as well as the bond lengths for the neutral and anionic five-membered rings investigated in this study are summarized in Table 2.

In general, B and N should be expected to form double bonds because of the electron deficient character of the B atom, which facilitates N lone pair donation into the empty 2p orbital of B. This expectation is consistent with the relative stabilities of the isomers of **C2N2B** and **C2NB2** and anions $C_xN_yB_zH_5^-$, since the most stable isomer has the most

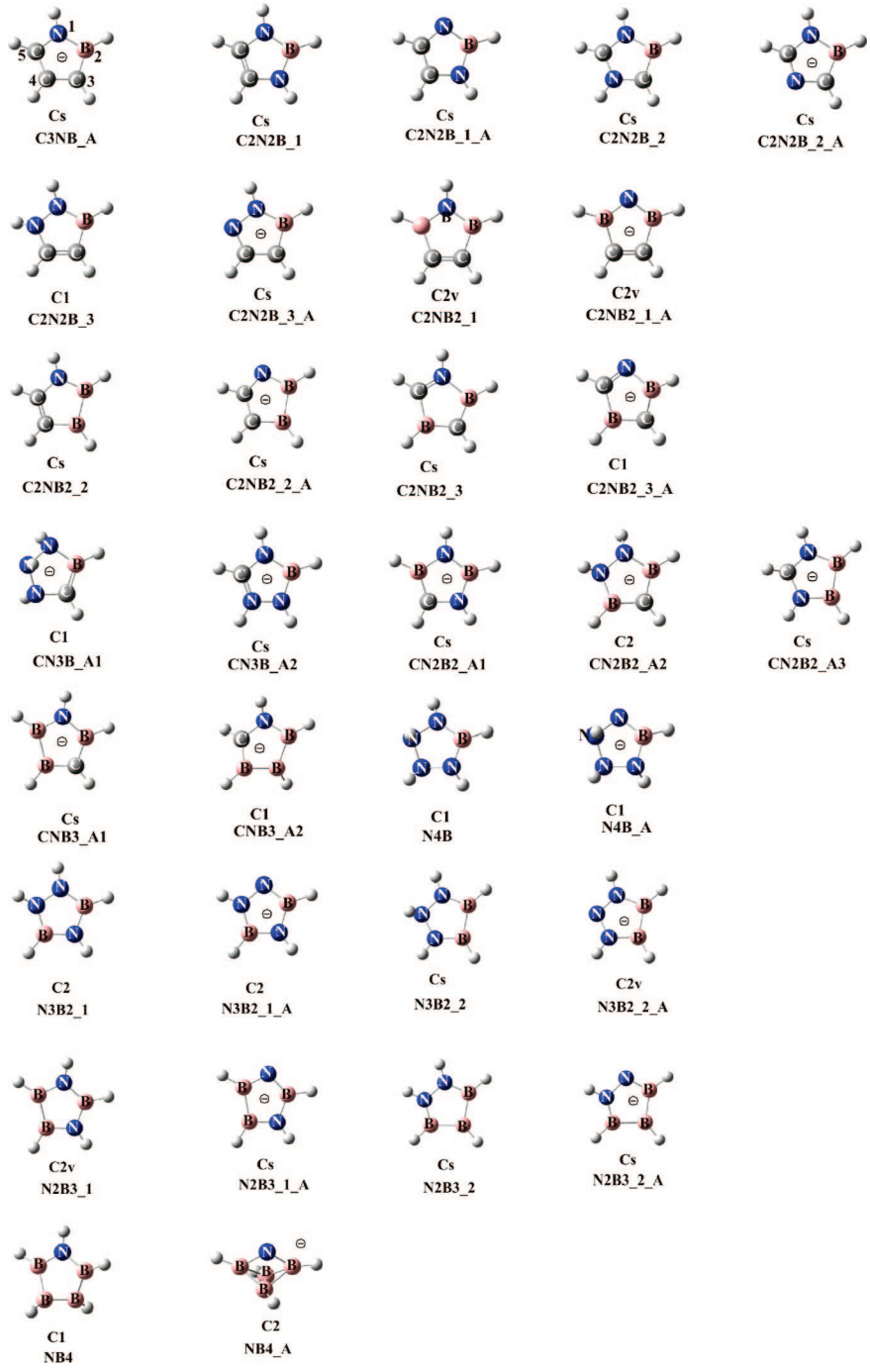



***Figure 1.*** Five-member BN-containing rings with the numbering system illustrated in **C3NB_A**.

BN bonds. Nevertheless, in most systems the BN bonds have BOs close to those in borazine, indicating a certain amount of electron delocalization.

As we shall discuss later in more detail, the NBO analysis reveals that some systems exhibit a clear electron delocalization through the participation of three-center bonds. Some others present what might appear to be more localized BN double bonds, but interactions involving these bonds with adjacent atoms in the rings result in BN bonds that still resemble those of borazine, with BOs close to 1.0 rather than 1.26. Why are the bond orders less than expected even though there exist both $\sigma$ and $\pi$ bonding components? This is a consequence of the extreme polarity of the BN bond. A particular BN bond can be either an electron donor since N is electron rich or an electron acceptor because B is electron poor. In the first case charge transfer can occur from a BN bonding orbital to an electron acceptor within the ring, normally a C or a B atom, and this leads to a decrease in the population of this bonding orbital and a decrease in the BO. In the second case, there is charge donation from an N electron donor within the ring into a BN antibonding orbital. The increased population of this antibonding orbital also leads to a decrease of the BO. Only when these interactions are weak does the BO approach the BO of $H_2N{=}BH_2$.

As mentioned above, several systems are characterized by the presence of three-center bonds. A three-center CNB bond is found in **C3NB_A** and **C2N2B_2** and its anion; NBN in **C2N2B_1** and its anion and **CN2B2_A1**; BCB in **C2NB2_3**

***Table 2.*** Relative Enthalpies (ΔH°, kJ mol$^{-1}$),[a] B−N and B−H Distances (R, Å), Wiberg Bonds Orders (BO), and Coupling Constants (FC, Hz) for Neutral and Anionic Five-Membered BN-Containing Rings

| | neutral | | | | | anion | | | |
|---|---|---|---|---|---|---|---|---|---|
| species | bond | ΔH° | R | BO | FC | ΔH° | R | BO | FC |
| **C3NB_A** | N1−B2 | | | | | | 1.470 | 0.99 | −25.7 |
| **C2N2B_1** | N1−B2 | 0.0 | 1.433 | 1.06 | −32.4 | | 1.426 | 1.27 | −19.5 |
| | B2−N3 | | 1.433 | 1.06 | −32.4 | | 1.465 | 0.98 | −24.1 |
| **C2N2B_2** | N1−B2 | 123 | 1.468 | 0.97 | −22.5 | | 1.473 | 0.96 | −25.3 |
| **C2N2B_3** | N1−B2 | 182 | 1.427 | 1.10 | −28.7 | | 1.452 | 1.06 | −27.3 |
| **C2NB2_1** | N1−B2 | 0.0 | 1.438 | 0.99 | −25.8 | | 1.416 | 1.24 | −15.6 |
| **C2NB2_2** | N1−B2 | 181 | 1.422 | 1.05 | −24.8 | | 1.446 | 0.95 | −12.7 |
| **C2NB2_3** | N1−B2 | 199 | 1.609 | 0.63 | −10.4 | | 1.548 | 0.80 | −7.9 |
| **CN3B-A1** | N1−B2 | | | | | | 1.528 | 0.84 | −15.0 |
| **CN2B2_A1** | N1−B2 | | | | | 0.0 | 1.437 | 1.07 | −31.8 |
| | B2−N3 | | | | | | 1.430 | 1.10 | −35.4 |
| | N1−B5 | | | | | | 1.494 | 0.91 | −21.2 |
| **CN2B2_A2** | N1−B2 | | | | | 128 | 1.473 | 0.96 | −22.5 |
| **CN2B2_A3** | N1−B2 | | | | | 174 | 1.472 | 1.03 | −18.4 |
| **CNB3_A1** | N1−B2 | | | | | 0.0 | 1.494 | 0.83 | −23.0 |
| | N1−B5 | | | | | | 1.430 | 1.07 | −21.3 |
| **CNB3_A2** | N1−B2 | | | | | 232 | 1.408 | 1.16 | −25.1 |
| **N4B** | N1−B2 | | 1.414 | 1.09 | −34.8 | | 1.386 | 1.50 | −19.0 |
| | B2−N3 | | 1.444 | 1.02 | −27.0 | | 1.514 | 0.84 | −14.6 |
| **N3B2_1** | N1−B2 | 0.0 | 1.424 | 1.10 | −32.7 | | 1.411 | 1.41 | −20.9 |
| | B2−N3 | | 1.440 | 1.02 | −28.7 | | 1.484 | 0.91 | −20.5 |
| | N3−B4 | | 1.440 | 1.02 | −28.7 | | 1.438 | 1.05 | −32.1 |
| | B4−N5 | | 1.424 | 1.10 | −32.7 | | 1.421 | 1.15 | −35.7 |
| **N3B2_2** | N1−B2 | 341 | 1.407 | 1.19 | −25.9 | | 1.467 | 1.09 | −19.2 |
| **N2B3_1** | N1−B2 | 0.0 | 1.451 | 0.94 | −29.6 | | 1.418 | 1.19 | −17.0 |
| | B2−N3 | | 1.451 | 0.94 | −29.6 | | 1.509 | 0.79 | −19.2 |
| | N3−B4 | | 1.430 | 1.07 | −20.5 | | 1.422 | 1.13 | −22.2 |
| | N1−B5 | | 1.430 | 1.07 | −20.5 | | 1.419 | 1.30 | −14.6 |
| **N2B3_2** | N1−B2 | 410 | 1.400 | 1.23 | −29.7 | | 1.403 | 1.32 | −18.5 |
| | B4−N5 | | 1.400 | 1.23 | −29.7 | | 1.395 | 1.55 | −29.8 |
| **NB4**[b] | N1−B2 | | 1.446 | 1.01 | −18.2 | | 1.474 | 0.96 | −3.0 |

| | | neutral | | anion | |
|---|---|---|---|---|---|
| species | bond | R | FC | R | FC |
| **C3NB_A** | B2−H | | | 1.207 | 108.3 |
| **C2N2B_1** | B2−H | 1.185 | 145.3 | 1.207 | 120.6 |
| **C2N2B_2** | B2−H | 1.187 | 137.1 | 1.204 | 112.2 |
| **C2N2B_3** | B2−H | 1.188 | 134.0 | 1.205 | 110.8 |
| **C2NB2_1** | B2−H | 1.192 | 128.9 | 1.221 | 108.9 |
| **C2NB2_2** | B2−H | 1.191 | 126.8 | 1.215 | 106.0 |
| | B3−H | 1.193 | 122.6 | 1.211 | 101.7 |
| **C2NB2_3** | B2−H | 1.192 | 136.6 | 1.213 | 109.5 |
| | B4−H | 1.194 | 128.1 | 1.215 | 103.4 |
| **CN3B-A1** | B2−H | | | 1.205 | 111.8 |
| **CN2B2_A1** | B2−H | | | 1.204 | 116.5 |
| | B5−H | | | 1.207 | 108.3 |
| **CN2B2_A2** | B2−H | | | 1.213 | 102.7 |
| **CN2B2_A3** | B2−H | | | 1.210 | 101.9 |
| **CNB3_A1** | B2−H | | | 1.218 | 98.7 |
| | B4−H | | | 1.218 | 95.1 |
| | B5−H | | | 1.218 | 94.6 |
| **CNB3_A2** | B2−H | | | 1.211 | 102.5 |
| | B3−H | | | 1.216 | 95.1 |
| | B4−H | | | 1.209 | 107.3 |
| **NB4** | B2−H | 1.184 | 145.6 | 1.205 | 121.5 |
| **N3B2_1** | B2−H | 1.187 | 141.4 | 1.206 | 118.7 |
| | B4−H | 1.187 | 141.4 | 1.202 | 118.5 |
| **N3B2_2** | B2−H | 1.190 | 126.7 | 1.207 | 104.9 |
| **N2B3_1** | B2−H | 1.191 | 134.4 | 1.219 | 113.3 |
| | B4−H | 1.194 | 122.7 | 1.216 | 97.8 |
| | B5−H | 1.194 | 122.7 | 1.224 | 104.5 |
| **N2B3_2** | B2−H | 1.192 | 126.9 | 1.213 | 107.3 |
| | B3−H | 1.192 | 119.7 | 1.211 | 96.3 |
| | B4−H | 1.192 | 126.9 | 1.208 | 106.5 |
| **NB4**[a] | B2−H | 1.192 | 132.1 | 1.190 | 160.5 |
| | B3−H | 1.189 | 137.0 | 1.191 | 158.8 |

[a] $\Delta H^\circ = \Delta E^\circ + \Delta ZPE$. [b] The structures of neutral and deprotonated **NB4** are very different. In the neutral ring, N1−B2 and N1−B5 are equivalent. In the anion, N1−B2 and N1−B4 are equivalent.

and its anion, **CN2B2_A2**, and **CNB3_A1**; and BNB in **C2NB2_1** and its anion. Thus, the similarity between the BNB bonding pattern for **C2NB2_1** and borazine, for instance, is nicely reflected in their corresponding ELF descriptions, as illustrated in Figure 2. Bonding in **NB4** involves three-center BBB bonds. The BOs for rings with three-center bonds are close to 1.0 as in borazine, but there are some exceptions. The most obvious is **C2NB2_3** and its

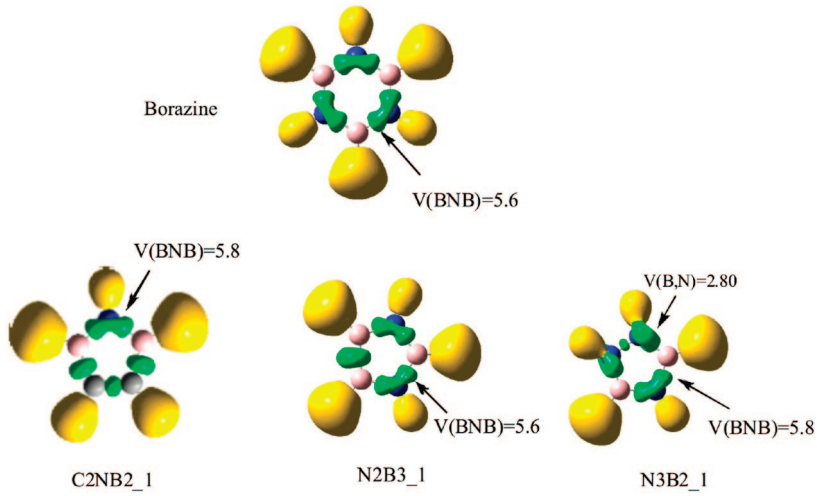



***Figure 2.*** Three-dimensional representations of ELF isosurfaces with ELF = 0.80 for borazine and selected five-membered BN-containing rings. Yellow lobes correspond to V(C,H) or V(N,H) basins. Green lobes correspond to V(C,C), V(C,N), V(B,N), and V(N,N) basins. Populations of V(B,N) basins are also indicated.

anion, both of which exhibit a three-center B2C3B4 bond and have N1 doubly bonded to C5. The existence of this double bond impedes conjugation between N1 and B2 with the result that the N1B2 bond in the neutral ring is the longest BN bond in the series and has the smallest BO of 0.63. B2−N1 conjugation increases slightly upon deprotonation of N1, as reflected by a BO of 0.80 in the anion, which, neverthless, still has the longest BN bond and the smallest BO among the anions.

In all the remaining cases in which the rings present one BN bond or two equivalent BN bonds, the NBO analysis gives two components, one $\sigma$ and one $\pi$. When the system presents two pairs of equivalent BN bonds, as in **N3B2_1**, two of the bonds appear as double bonds (N1B2 and N5B4) and two (N3B2 and N3B4) as single bonds in the NBO analysis. As mentioned above, the tendency of B and N to form double bonds is due to the electron deficient character of the B atom, which facilitates N lone pair donation into the empty 2p orbital of B. These BN double bonds involve an electron rich N atom which makes a large contribution (about 80%) to the BN bonding molecular orbital. Nevertheless, although the BN bonds might appear to be typical BN double bonds as in $H_2BNH_2$, interactions with other atoms in the ring lead to a significant electron delocalization and to BOs similar to those of borazine. Another interesting example is found for **CN2B2_A3.** In this molecule there is a strong charge donation from the N1B2 and N4B3 $\pi$ bonding orbitals to C5, a good electron acceptor. This favors the conjugation of C5 with the neighboring N atoms and is reflected in a CN bond order of 1.2. This donation reduces the electron population of the BN $\pi$ orbital, and the BO of this bond approaches that of borazine. A similar change is observed for **N2B3_1** as charge is transferred from the $\pi$ B5N1 and B4N3 bonding orbitals into the B2 empty 2p orbital, which enhances its conjugation with N1 and N3. Hence, relative to a single bond, the N1B2 and B2N3 bonds acquire some double bond character, while the N3B4 and B5N1 bonds lose double bond character, and the BOs of both sets of bonds approach that of borazine. The ELF descriptions of all of the BN bonds in this system closely resemble that of borazine, as can be seen in Figure 2.

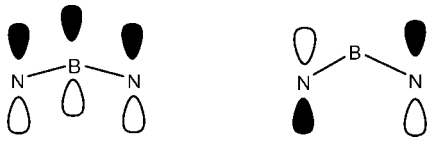
***Table 3.*** Intrinsic Acidities ($\Delta_{acid}H°$, kJ mol$^{-1}$) of the Neutral Five-Membered BN Rings

| molecule | $\Delta_{acid}H°$ |
|---|---|
| **C2N2B_1** | 1551 |
| **C2N2B_2** | 1471 |
| **C2N2B_3** | 1441 |
| **C2NB2_1** | 1556 |
| **C2NB2_2** | 1514 |
| **C2NB2_3** | 1527 |
| **N4B** | 1538 |
| **N3B2_1** | 1528 |
| **N3B2_2** | 1440 |
| **N2B3_1** | 1558 |
| **N2B3_2** | 1499 |
| **NB4** | 1417 |

A similar but much weaker effect is observed in **N2B3_2** because in this ring the acceptor boron atom B3 is bonded to two B atoms and has a natural charge of +0.19e (as compared to +0.81e for B2 in **N2B3_1**). Thus, electron donation from the BN bonds to B3 is minimal, and both BN bonds of **N2B3_2** exhibit double bond character with BOs close to 1.26, the BO of $H_2B{=}NH_2$. On the other hand, **N3B2_1** provides an example in which the N1B2 and N5B4 bonds act as electron acceptors through lone pair donation from the N3 lone-pair into $\pi_{BN}$* antibonding orbitals. This leads to a decrease of the BOs of the N1B2 and N5B4 bonds and an increase in the BOs of the N3B2 and N3B4 bonds. The net result is that the ELFs associated with the B2N3B4 fragment closely resemble that of borazine, as illustrated in Figure 2. A similar effect is found in **N3B2_2**, but it is significantly reduced since the electron donor is directly attached to the two electron-rich N atoms.

As expected, the most acidic site of a neutral BN ring is a NH group. The calculated intrinsic acidities, defined as the enthalpy at 0 K of reaction 1, are summarized in Table 3. All values are typically in the 1440−1560 kJ mol$^{-1}$ range, with **NB4** the only exception, most likely a reflection of the

***Scheme 1***




dramatic structural rearrangement that takes place upon deprotonation, as discussed below.

$$HA(g) \rightarrow H^+(g) + A^-(g) \quad (1)$$

The question to be addressed at this point is how do the bonding patterns found for molecules change when the most acidic N−H proton is removed to form the anion. Two groups of molecules can be distinguished: those in which the deprotonated nitrogen atom participates in a BN bond and those in which the deprotonated N atom is not bonded to a B atom. Among the molecules in the first set are **C2N2B_1**, **N3B2_1**, **N2B3_1**, and **N2B3_2**. For these, the deprotonation process is accompanied by a lowering of symmetry as two BN bonds which are equivalent in the molecule are no longer equivalent in the anion. **C2N2B_1** serves as a typical example. The N1B2N3 bonds in this molecule are three-center bonds. The orbitals involved in bonding are illustrated in Scheme 1. The first is completely bonding, while the second is BN nonbonding and NN antibonding. In contrast, in the anion the B 2p orbital contributes to the second three-center N1B2N3 orbital which then becomes B2N3 antibonding and B2N1 bonding. Hence, upon deprotonation the B2N3 BO decreases from 1.06 to 0.98, whereas the B2N1 BO increases from 1.06 to 1.27.

Also included in the first set are the molecules **N4B** and **NB4**, both of which experience dramatic structural and bonding changes upon deprotonation. In both neutral molecules all of the BN bonds have BOs near 1.0. Deprotonation of **N4B** occurs at N1 and reduces conjugation with N3. As a result, the N1B2 BO increases, and the electron density at the BCP also increases dramatically from 0.196 au to 0.227 au, while the B2N3 BO decreases and the electron density at the BCP also decreases from 0.201 au to 0.165 au. In **NB4_A**, a reinforcement of the N1B2 and N1B5 bonds upon deprotonation might be expected, but since the molecule contains four electron deficient centers, electron delocalization involves all of the atoms and leads to significant distortion of the five-membered ring. The molecular graph for the resulting anion which is given in Figure 3 shows bond paths only for N1B2 and N1B4. There is also a slight reduction in BOs, consistent with bond lengths of 1.473 Å. The other two BN interactions, N1B5 and N1B3, are across much longer distances of 1.798 Å and are essentially nonbonding. B3B4 and B2B5 are bonded, but no BCPs are associated with B4B5 and B2B3 at distances of 1.859 Å, so there are no B−B bonds in these two cases

There is only one molecule, **C2NB2_1**, for which deprotonation of the N atom participating in the BN bond is not accompanied by a symmetry change. In this case N1 deprotonation enhances conjugation of the nitrogen atom with the two neighboring B atoms. The BN BOs increase, and the electron densities at the BCPs also increase from 0.195 au to 0.215 au.

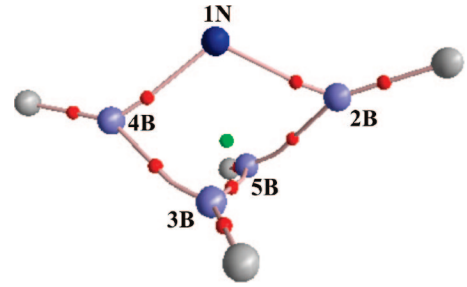



***Figure 3.*** Molecular graph of **NB4_A**. Red dots denote (3,-1) BCPs, and the green dot a (3,+3) cage critical point.

The second group of molecules are those in which the deprotonated nitrogen is not bonded to a B atom. The molecules **C2N2B_2** and **N3B2_2** and their anions illustrate two extremes in bonding changes. In **C2N2B_2** deprotonation of N4 alters the conjugation of N4 with the two C atoms directly bonded to it but has only a small effect on the remote N1B2 bond. The N1B2 BO for the anion is essentially identical to that for the neutral molecule. In contrast, deprotonation of N5 in **N3B2_2** has a significant but indirect effect on the BN bonds. In the anion, conjugation of N5 with the two N atoms directly bonded to it increases at the expense of reduced conjugation of N1 with B2 and N4 with B3. An increase in the BN bond distance and a decrease in BO are observed.

**Spin−Spin Coupling Constants.** Table 2 reports the FC terms for one-bond B−N and B−H coupling in the set of neutral and anionic five-membered rings. The values of $^1$J(B−N) vary considerably and are always negative. Since the magnetogyric ratio of $^{11}$B is positive and that of $^{15}$N is negative, the reduced coupling constant $^1$K(B−N) is positive and therefore in agreement with the Dirac vector model.[23] For the neutral rings, $^1$J(B−N) ranges from −10.4 Hz in **C2NB2_3** to −34.8 Hz in **N4B**, thereby bracketing the value of $^1$J(B−N) for borazine which is −28.7 Hz. $^1$J(B−N) spans an even greater range in the anions, from −3.0 to −35.7 Hz. It is interesting to note that the smallest absolute value of $^1$J(B−N) among neutral molecules is found for **C2NB2_3**, which has the weakest B−N bond. Among the anions, the smallest absolute value is found for **NB4_A**, which is structurally unique. Nevertheless, closer examination of $^1$J(B−N) for these systems indicates that there is no overall correlation between the value of this coupling constant and either the BN distance or the BN bond order. This is readily illustrated using the bond orders and B−N coupling constants for the N1−B2 bonds in **C2N2B_3** and **N3B2_1** as references. Both N1−B2 bonds have BOs of 1.10 and corresponding values of $^1$J(N1−B2) of −28.7 and −32.7 Hz, respectively. The N1−B2 bonds in **N3B2_2** and **C2N2B_2** have larger (1.19) and smaller (1.05) BOs, respectively, but the corresponding coupling constants are similar (−25.9 and −24.8 Hz) and reduced relative to the reference values. The lack of correlation between coupling constants and ground-state distances and bond orders is neither new nor surprising, since coupling constants are second-order properties which depend on electron distributions in both the ground state and the excited electronic states which couple to it primarily through the FC operator.

Are there any systematic trends among these coupling constants? The data of Table 2 indicate that the absolute value of $^1$J(B−N) decreases in the anion relative to the corresponding neutral ring if coupling involves the deprotonated N or a B atom bonded to a deprotonated N. However, if this is not the case, then $^1$J(B−N) may increase or decrease upon deprotonation. For example, $^1$J(N1−B2) for **C2N2B_3** and **N3B2_2** decrease in absolute value when N5 is deprotonated, while $^1$J(B4−N5) for **N3B2_1** and **N2B3_2** increase in absolute value when N1 is deprotonated.

Why does $^1$J(B−N) decrease when the coupled N is deprotonated or when the coupled B is bonded to a deprotonated N? Some insight into the answer to this question can be gained by noting that nitrogen deprotonation replaces an N−H bond with a nonbonding pair of electrons. The decrease in $^1$J(B−N) which is observed is consistent with the change in $^1$J(N−C) upon protonation of pyridine.[20] The experimental and computed EOM-CCSD values of $^1$J(N−C) for neutral pyridine are quite small at 0.67[24] and 0.53[20] Hz, respectively. The FC contribution to $^1$J(N−C) is −4.3 Hz and is counterbalanced by a PSO contribution of 5.5 Hz. In pyridinium, the absolute value of $^1$J(N−C) increases, with experimental and computed values of −12[24] and −15[20] Hz, respectively. In the ion, the FC term increases in absolute value to −18.3 Hz and dominates the PSO term which decreases slightly to 4.2 Hz. The change in the FC term for one-bond N−C coupling is associated with s-electron density differences in the ground and excited electronic states of pyridine and pyridinium. From a sum-overstates perspective, the particular excited $\sigma$-type states which dominate and determine the FC term are most probably quite different depending on the presence or absence of a heteroatom with a lone pair of electrons. Thus, the presence of nonbonding electrons plays an important role in determining both the sign and magnitude of the N−C coupling constant in pyridine and pyridinium. In the BN rings, the FC term is an excellent approximation to J, and the presence or absence of a lone pair on N has a major impact on $^1$J(B−N) when the coupled N is deprotonated or when the B is bonded to a deprotonated N. $^1$J(B−N) decreases when the N−H bond is replaced by a nitrogen nonbonding pair of electrons.

The B−H coupling constant $^1$J(B−H) is always positive and varies from 120 to 146 Hz in the neutral rings and from 95 to 119 Hz in the anions (excluding **NB4_A**). $^1$J(B−H) always decreases in the anion relative to the corresponding base, independent of whether or not the coupled B is bonded to the deprotonated N. The only exceptions are $^1$J(B2−H3) and $^1$J(B3−H4) for **NB4** which increase significantly upon deprotonation and, at approximately 160 Hz, are the largest B−H coupling constants found in the entire set of neutral and anionic rings. However, it should be noted that neutral and deprotonated **NB4** are significantly different structurally and that $^1$J(N1−B2) in the anion is only −3.0 Hz and therefore smaller than any other B−N coupling constant.

## Conclusions

In this study, the structural, bonding, and NMR spectroscopic properties of a series of five-membered neutral and anionic rings containing BN bonds $C_xN_yB_zH_5$ ($x + y + z = 5$), their most stable deprotonated anions $C_xN_yB_zH_4^-$, and anionic rings $C_xN_yB_zH_5^-$ have been examined. In the majority of these rings there is significant electron delocalization which leads to BOs close to that found in borazine which contains the prototypical aromatic BN bond. Only in a few systems are localized B=N double bonds found. Electron delocalization in these rings is favored by the strong polarity of the BN bonds, which behave either as electron donors or as electron acceptors depending on their bonding environment in the ring. Some significant bonding changes are found upon deprotonation of the neutral rings. In most cases these changes accompany a loss of symmetry which occurs with proton elimination.

Computed one-bond $^{15}$N−$^{11}$B spin−spin coupling constants range from −10.4 to −34.8 Hz in the neutral rings, thereby bracketing the value of $^1$J(B−N) for borazine, −28.7 Hz. $^1$J(B−N) spans an even greater range in the anions, from −3.0 to −35.7 Hz. The absolute value of $^1$J(B−N) decreases upon deprotonation if coupling involves the deprotonated nitrogen or a boron atom bonded to the deprotonated N. $^1$J(B−H) always decreases in the anion relative to the corresponding neutral ring.

**Acknowledgment.** This work has been partially supported by the DGI Project Nos. BQU2003-01251 and CTQ2006-08558/BQU and CONSOLIDER on Molecular Nanoscience CSD2007-00010 as well as by the Project MADRISOLAR. Ref.: S-0505/PPQ/0225 of the Comunidad Autónoma de Madrid and the COST Action. Thanks are due to the Ohio Supercomputer Center for continued support of this work.

## References

(1) Greenwood, N. N.; Earnshaw, A. *Chemistry of the Elements;* Pergamon Press: Oxford, 1984.

(2) Mó, O.; Yáñez, M.; Martín Pendás, A.; Del Bene, J. E.; Alkorta, I.; Elguero, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3970.

(3) Segawa, Y.; Yamashita, M.; Nozaki, K. *Science* **2006**, *314*, 113.

(4) Forster, T. D.; Krahulic, K. E.; Tuononen, H. M.; McDonald, R.; Parvez, M.; Roesler, R. *Angew. Chem., Int. Ed.* **2006**, *45*, 6356.

(5) Ly, H. V.; Forster, T. D.; Corrente, A. M.; Eisler, D. J.; Konu, J.; Parvez, M.; Roesler, R. *Organometallics* **2007**, *26*, 1750.

(6) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Phys. Chem. A* **2006**, *110*, 9959.

(7) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Chem. Theory Comput.* **2007**, *3*, 549.

(8) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Phys. Chem. A* **2007**, *111*, 419.

(9) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(10) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397.

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

(12) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597.

(13) Perera, S. A.; Sekino, H.; Bartlett, R. J. *J. Chem. Phys.* **1994**, *101*, 2186.

(14) Perera, S. A.; Bartlett, R. J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1995**, *117*, 8476.

(15) Perera, S. A.; Bartlett, R. J. *J. Am. Chem. Soc.* **1996**, *118*, 7849.

(16) Perera, S. A.; Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1996**, *104*, 3290.

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

(18) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572.

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

(20) Del Bene, J. E.; Elguero, J. *Magn. Reson. Chem.* **2006**, *44*, 784.

(21) Del Bene, J. E.; Alkorta, I.; Elguero, J. *J. Chem. Theory Comput.* **2008**, *4*, 967.

(22) Stanton, J. F. et al. *ACES II a program product of the Quantum Theory Project*; University of Florida: Gainesville, FL.

(23) Lynden-Bell, R. M.; Haris, R. K. *Nuclear Magnetic Resonance Spectroscopy*; Appleton Century Crofts: New York, 1969.

(24) Berger, S.; Braun, S.; Kalinowski, H.-O. *NMR Spectroscopy of the Non-Metallic Elements*; John Wiley & Sons: Chichester, 1996; p 267.

CT8002699

JCTC Journal of Chemical Theory and Computation

# Thermochemistry of Hydrocarbons. Back to Extended Hückel Theory

Alexander A. Voityuk*

*Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain, and Institute of Computational Chemistry, Universitat de Girona, 17071 Girona, Spain*

Received August 7, 2008

**Abstract:** A modified Extended Hückel method that provides accurate values of heats of formation and structural parameters of hydrocarbons is described. The results are reported for an extensive set of molecules and radicals belonging to different classes. The calculated heats of formation for 120 molecules and 26 radicals are close to the experimental data with the mean absolute error of 1.90 kcal/mol. The internal consistency of the calculated data allows reliable prediction of the reaction enthalpy for various hydrocarbon transformations. The proposed scheme is computationally very efficient, and the calculation of a large system requires only a few seconds on a PC. A computer program for the calculation is provided in the Supporting Information.

## 1. Introduction

The heat (or enthalpy) of formation $\Delta H^0_{f,298}$ of chemical compounds is a key thermodynamic parameter allowing for quantitative description of chemical reactions. Nowadays, experimental heats of formation are available for many hydrocarbons (see NIST Chemistry WEB book[1] and other compilations[2]); however, for molecules that include more than 6 carbon atoms, experimental data are often missing, especially if the compounds include unsaturated cyclic structures. For instance, experimental values of $\Delta H^0_{f,298}$ are available only for 6 isomers of $C_6H_6$,[1] though more than 200 chemically reasonable structures are possible. Because experimental measurements of accurate heats of formation are quite difficult, elaboration of computational methods to predict thermochemical data has attracted considerable attention. The simplest approach is based on the group additivity scheme;[3] however, this method can be applied only to certain types of molecules and radicals because for many groups the parameters have been undetermined.

Unlike the group additivity scheme, quantum chemical (QC) methods do not have such limitations. A dramatic increase in the computer power makes the QC techniques a good alternative to experiment. The developments and results in this area have been considered in several reviews.[4−6] Various QC approaches are used to derive $\Delta H^0_{f,298}$ and related data (atomization energies, thermochemical stability of isomers, bond dissociation energies, etc.). The composite *ab initio* schemes G2, G3, and G4 and their variants provide thermochemical data to chemical accuracy (±1 kcal/mol);[7−9] however, such calculations can be performed only for relatively small systems containing up to 15 heavy atoms.

Density functional theory (DFT) has become the method of choice for computational studies of medium-sized molecules.[10] The B3LYP scheme appears to be the most popular approach to calculate the energetics of chemical reactions; however, recent critical examinations of this and other popular functionals have shown that errors in heats of formation and reaction energies increase significantly with the size of systems.[11−14] For instance, B3LYP performs poorly in predicting the reaction energy of C−C and C−H bond splitting.[15,16] Probably, many difficulties can be overcome by using new exchange-correlation functionals.[17]

The *ab initio* and DFT techniques become inapplicable for extended hydrocarbons because of huge computational demands. QC calculations of large molecular systems can be carried out by using more efficient semiempirical methods.[18] New models[19−21] based on the neglect diatomic differential overlap (NDDO) scheme are enabled to obtain accurate data on the ground state energetics and structure. In many cases, the semiempirical estimates are of the same quality as the results of DFT calculations.[22]

* Corresponding author e-mail:alexander.voityuk@icrea.es.

10.1021/ct8003222 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/10/2008

***Table 1.*** Parameters of the MTB/2 Model

| Atomic Parameters | | |
|---|---|---|
| parameter | H | C |
| $U_s^A$ (eV) | −13.605 | −21.559 |
| $U_p^A$ (eV) | | −13.507 |

| Bond-Type Parameters | | | |
|---|---|---|---|
| parameter | H−H | C−H | C−C |
| $\lambda ss$ | 0.280 | 0.275 | 0.086 |
| $\lambda sp$ | | 0.218 | 0.180 |
| $\lambda pp\sigma$ | | | 0.186 |
| $\lambda pp\pi$ | | | 0.282 |
| $\beta ss$ (eV) | −4.442 | −8.574 | −5.969 |
| $\beta sp$ (eV) | | −6.813 | −6.160 |
| $\beta pp\sigma$ (eV) | | | −8.420 |
| $\beta pp\pi$ (eV) | | | −7.403 |
| $\alpha_{AB}$ (Å$^{-1}$) | 2.823 | 2.831 | 3.401 |
| $\gamma_{AB}$ (eV) | 12.612 | 99.370 | 658.659 |
| $\omega_{AB}$ (eV) | −0.0791 | −0.0340 | 0.0312 |
| $r_{AB}$ (Å) | 2.279 | 2.843 | 3.044 |

Another computationally efficient approach is based on a tight-binding (TB) model,[23,24] which bears close similarity to the extended Hückel method (EHM). Almost 50 years ago in his seminal work, Hoffmann employed the EHM method to calculate the structure and energetics of hydrocarbons.[25] While EHM was displaced from computational organic chemistry by NDDO-based semiempirical methods,[18] the TB schemes have been widely employed in computational physics.[23,24,26] The most accurate TB model is based on DFT.[24,27,28] Its performance in predicting the structure and energetics of organic molecules is comparable to that of the conventional semiempirical methods.[29−31]

Several TB schemes have been proposed for hydrocarbons;[32−40] however, most of them are not accurate enough for thermochemical calculations. The mean absolute error in $\Delta H^0_{f,298}$ and atomization energies is often essentially larger than 10 kcal /mol. In addition, the calculated structural parameters of hydrocarbons are less accurate than the results of semiempirical calculations (MNDO, AM1, PM3, and their variants). Recently, we developed two TB schemes that provide more accurate estimates of $\Delta H^0_{f,298}$ and molecular geometries for hydrocarbons.[40,41] The first scheme[40] is based on the conventional extended Hückel method and explicitly includes the overlap matrix of basis atomic orbitals (AOs). In the second approach,[41] the AOs are assumed to be orthogonal, and the overlap matrix is replaced by the unit matrix. This *orthogonal* model is simpler and computationally more efficient than EHM. While TB methods that employ nonorthogonal AOs are commonly considered to be more accurate and transferable,[23,24] the opposite has been found by Jasper et al. for aluminium clusters.[42] We have shown that also for hydrocarbons the *orthogonal* model MTB[41] performs better than the nonorthogonal scheme.[40] In particular, the mean absolute error (MAE) in $\Delta H^0_{f,298}$ calculated with MTB is only half as much as that of the *nonorthogonal model*.[41]

The purpose of the present work is to describe a new variant of the MTB,[41] hereafter referred to as MTB/2, and to demonstrate its performance. Like its predecessor, the MTB/2 method is based on the orthogonal model and thus is even simpler than EHM; however, the performance of MTB/2 is impressive. The calculated MAE of 1.90 kcal/mol in $\Delta H^0_{f,298}$ for 120 molecules and 26 radicals is considerably smaller than MAEs of the standard semiempirical methods AM1 and PM3 (13.0 and 8.2 kcal/mol, respectively). Notice that MTB/2 is at least an order of magnitude faster than AM1 and PM3. Like the previous schemes,[41] MTB/2 provides quite accurate geometries of hydrocarbons, and the MAEs in bond lengths and bond angles are 0.013 Å and 1.3°. Therefore, we believe that MTB/2 may be helpful for fast estimation of $\Delta H^0_{f,298}$, reaction energetics, and structural parameters for hydrocarbons.

***Table 2.*** Experimental and Calculated $\Delta H^0_{f,298}$ for $H_2$ and Alkanes[a]

| molecule | exp. | MTB/2 | Δ |
|---|---|---|---|
| hydrogen | 0.0 | 0.0 | 0.0 |
| methane | −17.8 | −14.4 | 3.4 |
| ethane | −20.0 | −18.2 | 1.8 |
| propane | −25.0 | −23.9 | 1.1 |
| butane | −30.1 | −28.9 | 1.2 |
| pentane | −35.1 | −33.9 | 1.2 |
| n-hexane | −39.9 | −38.9 | 1.0 |
| heptane | −44.9 | −43.8 | 1.1 |
| octane | −49.8 | −48.8 | 1.0 |
| nonane | −54.5 | −53.8 | 0.7 |
| decane | −59.8 | −58.7 | 1.1 |
| undecane | −64.8 | −63.7 | 1.1 |
| dodecane | −69.2 | −68.6 | 0.6 |
| tridecane | −74.5 | −73.6 | 0.9 |
| tetradecane | −79.4 | −78.6 | 0.8 |
| pentadecane | −84.8 | −83.5 | 1.3 |
| hexadecane | −89.6 | −88.5 | 1.1 |
| 2-methylpropane | −32.3 | −30.7 | 1.6 |
| 2-methylbutane | −36.7 | −35.9 | 0.8 |
| 2,2-dimethylpropane | −40.1 | −38.1 | 2.0 |
| 3-methylpentane | −41.1 | −39.8 | 1.3 |
| 2,3-dimethylbutane | −42.0 | −41.7 | 0.3 |
| 2,2-dimethylbutane | −44.0 | −43.0 | 1.0 |
| 2-methylpentane | −41.7 | −40.9 | 0.8 |
| 3-ethylpentane | −45.4 | −47.8 | −2.4 |
| 3-methylhexane | −45.7 | −44.9 | 0.8 |
| 3,3-dimethylpentane | −48.2 | −48.3 | −0.1 |
| 2-methylhexane | −46.6 | −45.9 | 0.7 |
| 2,2-dimethylpentane | −49.3 | −47.9 | 1.4 |
| 2,4-dimethylpentane | −48.3 | −48.0 | 0.3 |
| 2,2,3-trimethylbutane | −48.9 | −49.6 | −0.7 |
| 3,3-dimethylhexane | −52.6 | −53.1 | −0.5 |
| 3,4-dimethylhexane | −50.9 | −51.4 | −0.5 |
| 3-ethyl-2-methylpentane | −50.4 | −53.3 | −2.9 |
| 2,2,4-trimethylpentane | −53.5 | −53.0 | 0.5 |
| 2,3,3-trimethylpentane | −51.6 | −54.0 | −2.4 |
| 2,3,4-trimethylpentane | −52.0 | −54.1 | −2.1 |
| 2,5-dimethylhexane | −53.2 | −53.5 | −0.3 |
| 2,3-dimethylhexane | −51.1 | −52.3 | −1.2 |
| 2,4-dimethylhexane | −52.3 | −50.1 | 2.2 |
| 2,2-dimethylhexane | −53.7 | −52.8 | 0.9 |
| 2,2,3,3-tetramethylbutane | −54.0 | −57.3 | −3.3 |
| 2,2,3-trimethylpentane | −52.6 | −54.9 | −2.3 |
| 3,3-diethylpentane | −55.5 | −56.1 | −0.6 |
| 2,2,4,4-tetramethylpentane | −57.7 | −57.9 | −0.2 |
| 2,2,3,3-tetramethylpentane | −56.7 | −61.4 | −4.7 |
| tri-t-butylmethane | −47.2 | −56.2 | 9.0 |
| mean error (47 comparisons) | | | 0.40 |
| mean abs. error | | | 1.42 |
| standard deviation | | | 2.06 |

[a] In kcal/mol.

***Table 3.*** Experimental and Calculated $\Delta H^0_{f,298}$ for Unsaturated Aliphatic Hydrocarbons[a]

| molecule | MTB/2 | exp. | Δ |
|---|---|---|---|
| ethylene | 15.7 | 12.5 | 3.2 |
| acetylene | 53.5 | 54.5 | −1.0 |
| propene | 6.7 | 4.9 | 1.8 |
| propyne | 45.2 | 44.2 | 1.0 |
| allene | 46.8 | 45.5 | 1.3 |
| 1-butene | 2.8 | 0.0 | 2.8 |
| trans-2-butene | −1.6 | −2.7 | 1.1 |
| cis-2-butene | −2.2 | −1.7 | −0.5 |
| isobutene | −2.7 | −4.0 | 1.3 |
| 1,2-butadiene | 38.7 | 38.8 | −0.1 |
| 1,3-trans-butadiene | 28.1 | 26.3 | 1.8 |
| 1-butyne | 39.9 | 39.5 | 0.4 |
| 2-butyne | 37.1 | 34.8 | 2.3 |
| vinylacetylene | 67.2 | 70.4 | −3.2 |
| diacetylene | 107.1 | 111.0 | −3.9 |
| trimethylethene | −10.2 | −9.9 | −0.3 |
| tetramethylethene | −19.4 | −16.8 | −2.6 |
| trans-1,3-pentadiene | 19.8 | 18.2 | 1.6 |
| cis-1,3-pentadiene | 20.0 | 19.5 | 0.5 |
| isoprene | 19.0 | 18.0 | 1.0 |
| 1,4-pentadiene | 29.4 | 25.2 | 4.2 |
| mean error (21 comparisons) | | | 0.59 |
| mean abs. error | | | 1.71 |
| standard deviation | | | 2.13 |

[a] In kcal/mol.

The paper is organized as follows. In the next section, the method is outlined. Then the performance of the proposed model is considered by comparing MTB/2 estimates with the experimental data and results of G3 and DFT calculations. Finally, conclusions are given and a further possible elaboration of the model is noted. A computer program for MTB/2 calculations is provided in the Supporting Information.

## 2. Method

The total energy of the system is calculated as a sum of the electronic energy and short-range repulsion energy:

$$E = \sum_i n_i \varepsilon_i + E_{rep} \tag{1}$$

In eq 1, $\varepsilon_i$ is an eigenvalue of the effective one-electron Hamiltonian $H$, $\varepsilon_i = \langle \psi_i | H | \psi_i \rangle$, and $n_i$ is the occupation number of MO $\psi_i$. The repulsion energy $E_{rep}$ is approximated by a sum of two-center potentials $G_{AB}$ depending only on the distance between atoms $A$ and $B$:

$$E_{rep} = \sum_{A>B} G_{AB} \tag{2}$$

The potential $G_{AB}$ includes implicitly terms of different physical nature (the core-core repulsion, correction due to double counting of the two-electron interactions, the exchange and correlation energy). It is determined by fitting to reference data. Accounting for $E_{rep}$ suggested by Chadi in 1979[43] is significant to get an accurate and transferable TB model.[23,24] Note that much earlier Hoffmann discussed the inclusion of such a term into the EHM total energy.[25]

In the general case (nonorthogonal model), the orbital energies $\varepsilon_i$ are obtained by solving the eigenvalue equations

$$\sum_i (H_{\mu\nu} - \varepsilon_i S_{\mu\nu}) c_{\nu i} = 0 \tag{3}$$

***Table 4.*** Experimental and Calculated $\Delta H^0_{f,298}$ for Cyclic Compounds[a]

| molecule | MTB/2 | exp. | Δ |
|---|---|---|---|
| cyclopropane | 16.7 | 12.7 | 4.0 |
| cis-dimethylcyclopropane | −0.7 | 1.3 | −2.0 |
| cyclopropene | 62.5 | 66.2 | −3.7 |
| 1-methylcyclopropene | 53.4 | 58.2 | −4.8 |
| 1,2-dimethylcyclopropene | 44.2 | 46.4 | −2.2 |
| methylenecyclopropane | 46.5 | 47.9 | −1.4 |
| cyclobutane | 5.2 | 6.8 | −1.6 |
| cyclobutene | 43.3 | 37.5 | 5.8 |
| 1,2-dimethylcyclobutene | 24.3 | 19.8 | 4.5 |
| methylenecyclobutane | 28.4 | 29.0 | −0.6 |
| cyclopentane | −20.6 | −18.3 | −2.3 |
| cyclopentene | 7.3 | 8.1 | −0.8 |
| cyclopentadiene | 34.2 | 32.1 | 2.1 |
| 3,4−dimethylenecyclobutene | 83.7 | 80.4 | 3.3 |
| fulvene | 51.1 | 53.5 | −2.4 |
| cyclohexane, chair | −29.2 | −29.5 | 0.3 |
| cyclohexene, half-chair | −2.1 | −1.2 | −0.9 |
| 1,3-cyclohexadiene | 20.6 | 25.0 | −4.4 |
| 1,4-cyclohexadiene | 23.0 | 24.0 | −1.0 |
| cycloheptatriene | 42.3 | 43.2 | −0.9 |
| cyclooctatetraene | 66.9 | 70.7 | −3.8 |
| bicyclobutane | 46.6 | 51.9 | −5.3 |
| trans-bicyclopropyl | 33.9 | 30.9 | 3.0 |
| bicycle[2.1.0]pentane | 38.5 | 37.8 | 0.7 |
| spirocyclopentane | 46.8 | 44.3 | 2.5 |
| norbornane | −5.4 | −13.1 | 7.7 |
| norbornadiene | 58.3 | 58.8 | −0.5 |
| cyclo[2,2,2]-octane | −23.7 | −23.6 | −0.1 |
| adamantane | −28.1 | −32.2 | 4.1 |
| cubane | 149.8 | 148.7 | 1.1 |
| mean error (30 comparisons) | | | 0.01 |
| mean abs. error | | | 2.59 |
| standard deviation | | | 3.24 |

[a] In kcal/mol.

As already mentioned, the scheme becomes faster and probably more accurate if an orthogonal AO basis is employed (it means that the overlap integrals in eq 3 are set to zero, $S_{\mu\nu} = \delta_{\mu\nu}$).[41] The MTB/2 method is based on the orthogonal scheme $(H-\varepsilon)\,C = 0$. A minimal valence basis of atomic orbitals (one 1s AO for H and four AOs 2s, $2p_x$, $2p_y$, and $2p_z$ for C) is used. The basis set is not explicitly constructed but just has the same symmetry properties as the atomic orbitals. The Hamiltonian operator is replaced with a matrix $H$, which is defined through semiempirical parameters.

In MTB/2, the diagonal matrix elements $H_{\mu\mu}$ is determined by the parameter $U_\mu$, corresponding to the energy of an electron in AO $\phi_\mu$ ($\mu = s$ and $p$); while off-diagonal one-center matrix elements are set to zero:

$$H_{\mu\nu} = U_\mu(A) \cdot \delta_{\mu\nu}(\mu, \nu \in A) \tag{4}$$

Two-center matrix elements (in the diatomic coordinate system) are defined using an empirical formula suggested by Kolb and Thiel[44]

$$H_{\mu\nu} = \pm \beta^{AB}_{\mu\nu} (R_{AB}/a_0)^{1/2} \exp\left(-\lambda^{AB}_{\mu\nu} R^2_{AB}/a_0^2\right) \tag{5}$$

where $\beta^{AB}_{\mu\nu}$ and $\lambda^{AB}_{\mu\nu}$ ($\mu \in A$, $\nu \in B$) are adjustable parameters. A suitable phase factor $\pm 1$ has to be used to retain the rotational invariance of the method. $R_{AB}$ is the interatomic distance in Å, and the Bohr radius $a_0$=0.52917 Å. As the Hamiltonian does not explicitly depend on the density matrix,

***Table 5.*** Experimental and Calculated $\Delta H^0_{f,298}$ for Aromatic Compounds and Fullerenes[a]

| molecule | MTB/2 | exp. | Δ |
|---|---|---|---|
| benzene | 20.5 | 19.7 | 0.8 |
| toluene | 12.2 | 12.0 | 0.2 |
| ethylbenzene | 8.3 | 7.1 | 1.2 |
| styrene | 34.9 | 35.3 | −0.4 |
| indene | 40.4 | 39.1 | 1.3 |
| mesitylene | −3.4 | −3.8 | 0.4 |
| naphthalene | 35.0 | 35.9 | −0.9 |
| anthracene | 52.2 | 55.2 | −3.0 |
| phenanthrene | 49.3 | 49.0 | 0.3 |
| azulene | 66.0 | 69.1 | −3.1 |
| biphenylene | 98.2 | 100.5 | −2.3 |
| pyrene | 54.3 | 53.9 | 0.4 |
| triphenylene | 69.2 | 65.5 | 3.7 |
| benzo[c]phenanthrene | 69.6 | 69.6 | 0.0 |
| benzo[a]anthracene | 66.1 | 68.9 | −2.8 |
| chrysene | 65.5 | 63.0 | 2.5 |
| naphthacene | 70.3 | 73.3 | −3.0 |
| perylene | 73.4 | 76.3 | −2.9 |
| coronene | 75.3 | 73.5 | 1.8 |
| corannulene | 110.8 | 110.1 | 0.7 |
| fullerene $C_{60}$ | 603.9 | 604.8 | −0.9 |
| fullerene $C_{70}$ | 653.1 | 653.8 | −0.7 |
| mean error (22 comparisons) | | | −0.30 |
| mean abs. error | | | 1.52 |
| standard deviation | | | 1.94 |

[a] In kcal/mol.

***Table 6.*** Experimental and Calculated $\Delta H^0_{f,298}$ for Radicals[a]

| radical | MTB/2 | exp. | deviation |
|---|---|---|---|
| methyl | 39.8 | 35.0 | 4.8 |
| ethyl | 29.7 | 28.9 | 0.8 |
| vinyl | 67.5 | 64.0 | 3.5 |
| ethynyl | 113.0 | 130.0 | −17.0 |
| 1-propyl | 23.3 | 23.3 | 0.0 |
| i-propyl | 18.3 | 19.1 | −0.8 |
| allyl | 40.0 | 39.3 | 0.7 |
| propargyl | 82.1 | 82.0 | 0.1 |
| cyclopropyl | 64.4 | 66.9 | −2.5 |
| cyclopropenyl | 102.7 | 105.0 | −2.3 |
| n-butyl | 18.4 | 19.1 | −0.7 |
| s-butyl | 13.9 | 17.0 | −3.1 |
| t-butyl | 7.4 | 9.5 | −2.1 |
| cyclobutyl | 50.2 | 51.1 | −0.9 |
| 1-pentyl | 13.4 | 10.1 | 3.3 |
| 2-pentyl | 8.0 | 7.4 | 0.6 |
| 2-me,2-butyl | 1.0 | 3.1 | −2.1 |
| neo-pentyl | 9.6 | 10.1 | −0.5 |
| 1,4-pentadiene-3-yl | 47.5 | 49.8 | −2.3 |
| cyclopentyl | 20.6 | 25.6 | −5.0 |
| spiropentyl | 90.8 | 90.9 | −0.1 |
| cyclopentadienyl | 61.7 | 58.0 | 3.8 |
| 2-hexyl | 3.0 | 3.0 | 0.0 |
| cyclohexyl | 12.2 | 13.9 | −1.7 |
| phenyl | 72.5 | 75.8 | −3.3 |
| benzyl | 46.5 | 49.0 | −2.5 |
| mean error (26 comparisons) | | | −1.12 |
| mean abs. error | | | 2.47 |
| standard deviation | | | 4.16 |

[a] In kcal/mol.

only a single diagonalization of the $H$ matrix is required to calculate the total energy; also, there is no difference by the treatment of closed- and open-shell systems.

The short-range potential $G_{AB}$ consists of two terms:

$$G_{AB} = \gamma_{AB}\exp(-\alpha_{AB}R_{AB}) + \omega_{AB}\exp[-6(R_{AB} - r_{AB})^2] \quad (6)$$

The first term approximates the repulsion potential between atoms $A$ and $B$, whereas the last one corrects the repulsion at distances close to the van der Waals contact $r_{AB}$.

The standard heat of formation at 298 K, $\Delta H^0_{f,298}$, is computed using the same procedure as employed in semiempirical methods

$$\Delta H^0_{f,298} = E - \sum_A E_{isol}(A) + \sum_A H^0_{f,298}(A) \quad (7)$$

where $E$ is the total energy of the molecule defined by eq 1, and $E_{isol}(A)$ is the energy of an isolated atom $A$

$$E_{isol}(A) = n_s U_s(A) + n_p U_p(A) \quad (8)$$

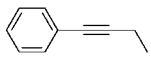
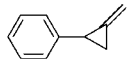
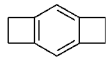
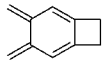
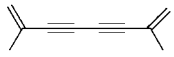
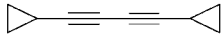
Here, $n_s$ and $n_p$ denote the number of valence $s$- and $p$-electrons in atom $A$. $\Delta H^0_{f,298}(A)$ is the experimental heat of formation of atom $A$, and for hydrogen and carbon $\Delta H^0_{f,298}(A)$ is 52.1 kcal/mol and 171.29 kcal/mol, respectively.[1] Note that in the previous version of the MTB method[41] as well as in several semiempirical programs an obsolete value of $\Delta H^0_{f,298}(C) = 170.89$ kcal/mol is employed to retain the consistency with previous results. Obviously, this difference of 0.4 kcal/mol does not affect calculated reaction enthalpies, e.g. the relative stability of isomers; however, the absolute values of $\Delta H^0_{f,298}$ of a system will change by $0.4\times N_C$ kcal/mol with the number of carbon atoms $N_C$ in the molecule. For compounds with $N_C \geq 10$, this difference is remarkable, and reparametrization of the computational scheme is required when using the new value of $\Delta H^0_{f,298}(C)$.

Notice that the zero-point vibrational energy and the enthalpy to heat the molecule from T=0 K to T=298 K are implicitly incorporated into the total energy $E$ through the parametrization (these terms are absorbed in some way into the parameters). Therefore, unlike *ab initio* and DFT, the calculations of the zero-point vibrational and heating energies are not required within MTB/2 to estimate $\Delta H^0_{f,298}$ and reaction enthalpies.

The parameters of the effective Hamiltonian and the repulsive potential $G_{AB}$ were derived as follows. The parameter $U_s$ for hydrogen was set to −13.605 eV (the negative of the ionization potential of the atom). Note that a shift of $U_s$ and $U_p$ parameters for H and C by a constant $\delta$ does not affect $\Delta H^0_{f,298}$ of neutral species, although it changes absolute values of orbital energies $\varepsilon_i \rightarrow \varepsilon_i + \delta$. All other parameters were fitted using experimental heats of formation[1,2] and structural parameters for a training set of molecules and radicals. Unlike the MTB scheme,[41] where the parameter $r_{AB}$ (eq 6) is determined as the sum of the van der Waals radii of atoms A and B, in MTB/2, $r_{AB}$ is an adjustable parameter.

The molecules and radicals in training sets were chosen to represent the most common bonding situations in hydrocarbons. Despite remarkable disparities among experimental heats of formation of $C_{60}$ (the measured values of $\Delta H^0_{f,298}$ range from 599.1 to 634.9 kcal/mol),[1,45] we included $C_{60}$ in the training set. As a reference value, we adopted $\Delta H^0_{f,298} =$ 604.6 kcal/mol employed by Green et al. for parameterization of a group additivity method for polycyclic aromatic hydrocarbons.[45] A nonlinear least-squares method was used to

***Table 7.*** Comparison of Experimental and Calculated Values of $\Delta H^0_{f,298}$ and Relative Energies of $C_6H_6$ and $C_{10}H_{10}$ Isomers[a]

| | $\Delta H^0_{f,298}$, in *kcal/mol* | | Relative energies, in *kcal/mol* | | | |
|---|---|---|---|---|---|---|
| $C_6H_6$ | Exp | MTB/2 | Exp | MTB/2 | B3LYP | MP2 |
|  | 19.8 | 21.3 | 0.0 | 0.0 | 0.0 | 0.0 |
|  | 53.6 | 51.3 | 33.8 | 30.0 | 34.7 | 36.0 |
|  | 80.4 | 82.6 | 60.6 | 61.4 | 62.9 | 67.9 |
|  | 90.2 | 91.8 | 70.4 | 70.5 | 68.9 | 69.5 |
|  | 94.6 | 93.1 | 74.8 | 71.8 | 82.1 | 91.7 |
| $C_{10}H_{10}$ | Exp | MTB/2 | Exp | MTB/2 | B3LYP | MP2 |
|  | 53.5 | 56.2 | 0.0 | 0.0 | 0.0 | 0.0 |
|  | 59.4 | 61.4 | 5.9 | 5.3 | -8.5 | 1.2 |
|  | 70.0 | 67.8 | 16.5 | 11.6 | 0.7 | 12.4 |
|  | 74.0 | 77.0 | 20.5 | 20.9 | 3.1 | 10.1 |
|  | 79.9 | 74.8 | 26.3 | 18.6 | 20.3 | 22.8 |
|  | 85.8 | 78.8 | 32.3 | 22.6 | 17.6 | 31.1 |
|  | 118.1 | 116.8 | 64.6 | 60.6 | 48.8 | 63.0 |
|  | 134.3.1 | 138.5 | 80.8 | 82.3 | 71.2 | 81.0 |

[a] B3LYP/6-31G** and MP2/6-31G** data were taken from the paper by Schreiner et al.[14]

optimize the parameters. Several parameterization runs were carried out starting from different parameter values and using different training sets. To choose the parameter set that yields most balanced results, extensive survey calculations were performed with several sets of parameters. The final MTB/2 parameters are listed in Table 1. A simple computer program for MTB/2 calculations of hydrocarbons under the MS Windows environment is provided in the Supporting Information.

## 3. Results and Discussion

**3.1. Heats of Formation.** Tables 2−6 contain the calculated and experimental values of $\Delta H^0_{f,298}$ for 120 molecules and 26 radicals that belong to different classes of hydrocarbons. The experimental data are taken from the standard sources[1,2] unless otherwise specified. The MAE of 1.90 kcal/mol is considerably smaller than those of the conventional semiempirical methods. For the same set of molecules, the

MAE of AM1 and PM3 is 13.0 and 8.2 kcal/mol, respectively. Excluding fullerenes $C_{60}$ and $C_{70}$ from the statistics (the $\Delta H^0_{f,298}$ values for these compounds are substantially overestimated by AM1 and PM3), one gets the MAEs to be 7.3 and 5.0 kcal/mol, which are still essentially larger than that of MTB/2.

For 16 medium-sized hydrocarbons of the G3 subset (see Table S1 in the Supporting Information), the MTB/2 MAE of 1.65 kcal/mol is considerably smaller than that of B3LYP (9.64 kcal/mol), AM1 (6.45 kcal/mol), and PM3 (3.44 kcal/mol). Various computational schemes of the G3 family provide very accurate results with the MAE ranging from 0.7 to 1.0 kcal/mol;[12] however, such estimations are much more time-consuming (by a factor of $10^5$) than the MTB/2 calculations.

*Alkanes.* As seen from Table 2, small deviations of the calculated values from the experimental $\Delta H^0_{f,298}$ are found both for short and long alkanes. The MAE for 47 molecules including normal and branched compounds is 1.42 kcal/mol. MTB/2 consistently overestimates $\Delta H^0_{f,298}$ n-alkanes $C_nH_{2n+2}$ by 0.40 kcal/mol. The $CH_2$ increment is well reproduced. Notice that this is not a common mark of QC methods. For instance, the error in the G3 enthalpies increases with chain length ranging from 0.25 to 1.93 kcal/mol.[16] The B3LYP error in $\Delta H^0_{f,298}$ amounts to −1.5 kcal/mol for propane and −30.3 kcal/mol for hexadecane. Note that the small effect of conformational averaging on the enthalpies of the n-alkanes can be neglected.[11]

The $\Delta H^0_{f,298}$ values of branched hydrocarbons are also well reproduced (Table 2). Thus, MTB/2 shows a good performance in estimating the energy differences between sterically crowded hydrocarbons and their linear isomers. For instance, the conversion enthalpy of neopentane to pentane is predicted to be 4.2 kcal/mol in good agreement with the experimental value of 5.1 kcal/mol. In contrast, B3LYP shows rather large branching errors (10−15 kcal/mol).[14] MTB/2 overestimates the stability of compounds with neighboring methyl groups, the deviation of −4.7 kcal/mol is found for 2,2,3,3-tetramethylpentane (Table 2), whereas $\Delta H^0_{f,298}$ of the 2,2,4,4-isomer is well reproduced (the error is only −0.2 kcal/mol).

*Unsaturated Aliphatic Hydrocarbons.* Table 3 compares the calculated and experimental heats of formation for 21 compounds with double and triple bonds. The MAE of MTB/2 is found to be 1.71 kcal/mol. On average, the scheme underestimates the stability of these compounds, and the mean deviation is 0.59 kcal/mol. The largest deviation of 4.18 kcal/mol is found for 1,4-pentadiene.

*Cyclic Hydrocarbons.* Analysis of the data for 30 cyclic molecules (Table 4) shows that the MAE for these species is somewhat larger, 2.59 kcal/mol. On average, MTB/2 overestimates the stability of 21 *monocyclic* molecules by 0.61 kcal/mol. The largest deviation of 5.80 kcal/mol is calculated for cyclobutene. Table 4 lists also heats of formation for nine *bi-* and *polycyclic* hydrocarbons. Such molecules are known to be quite difficult for semiempirical methods. MTB/2 gives surprisingly good estimates of $\Delta H^0_{f,298}$ for most compounds. A relatively large deviation is found for norbornane (7.7 kcal/mol).

*Aromatic Compounds and Fullerenes.* MTB/2 provides reliable heats of formation for 22 molecules listed in Table 5. The MAE is only 1.52 kcal/mol. The largest deviation with experiment of 3.7 kcal/mol is obtained for triphenylene. Experimental values of $\Delta H^0_{f,298}$ are available only for two fullerenes, $C_{60}$ and $C_{70}$. Large disparity among the experimental heats of formation has been already mentioned. For $\Delta H^0_{f,298}$ of $C_{60}$ and $C_{70}$, MTB/2 gives 604.8 and 653.1 kcal/mol which are in excellent agreement with the most reliable experimental values of 604.8 and 653.8 kcal/mol (Table 5).[45] Note that the standard and new semiempirical NDDO methods do not provide acceptable values of $\Delta H^0_{f,298}$ for fullerenes. For instance, the AM1 and PM3 errors in the heat of formation of $C_{70}$ are 407.3 and 229.5 kcal/mol, respectively.

*Radicals.* Even for open-shell systems, the scheme gives quite accurate estimates of $\Delta H^0_{f,298}$ (Table 6). The calculated heats of formation are in good agreement with experiment.[46] For 26 radicals, the MAE of MTB/2 is 2.47 kcal/mol (Table 6). The largest deviation of 17 kcal/mol is found for the ethynyl radical in which one carbon has the only adjacent atom. It suggests that MTB/2 is of limited use for treatment of R-C≡C$^\bullet$ radicals. If the ethynyl radical is excluded from statistics, we obtain the MAE=1.89 kcal/mol, which is similar to that found for closed-shell systems.

**3.2. Isomerization Reactions.** The ability of carbon atoms to form single, double, and triple bonds leads to the great variety of hydrocarbons. As the CC bond energy ranges from 80 kcal/mol for C−C to 200 kcal/mol for C≡C, the enthalpy of isomerization reactions (*i.e.* relative values of $\Delta H^0_{f,298}$ for isomers) may be very helpful for assessing the performance and inherent consistency of a computational method. In Table 7, we compare experimental and calculated heats of formation of $C_6H_6$ and $C_{10}H_{10}$ isomers. B3LYP and MP2 results for the isomers were recently reported by Schreiner et al.[14] Overall, the MTB/2 scheme provides the most accurate estimates. For $C_6H_6$, the three methods show a comparable accuracy; however, the stability of the last isomer is remarkably underestimated by B3LYP and MP2 (the errors are 7.3 and 16.9 kcal/mol), whereas MTB/2 gives an accurate estimate.

For $C_{10}H_{10}$ isomers, the MAE of MTB/2 is found to be 3.4 kcal/mol. The MP2 data are also in good agreement with experiment. In contrast, the relative energies of isomers derived from B3LYP seem to be unreliable. The data presented in Table 7 suggest that MTB/2 provides internally consistent estimates of $\Delta H^0_{f,298}$ across different types of hydrocarbons.

**3.3. [2+2] Cyclization of Tetramethylethene and Various Alkenes.** Table 8 compares the energy for the reaction calculated with different QC approaches. As reference data for the reaction energy, we use G3MP2 results.[16] G3MP2 predicts all the reactions to be exothermic, although the formation of crowded cyclobutanes has slightly smaller reaction energy. The MTB/2 estimates are in very good

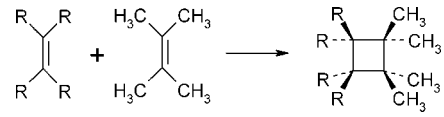


***Table 8.*** Predicted Energies (kcal/mol) for the [2+2] Cyclization of Tetramethylethene and Various Alkenes[a]

| alkene | G3MP2 | MP2 | B3LYP | MTB/2 |
|---|---|---|---|---|
| ethene | **−19.3** | −27.2 | −7.2 | −20.1 |
| propene | **−17.4** | −26.9 | −4.4 | −18.6 |
| isobutene | **−15.7** | −23.4 | 2.2 | −16.1 |
| *Z*-butene | **−18.0** | −25.8 | −0.1 | −17.7 |
| *E*-butene | **−16.7** | −27.7 | −2.0 | −17.3 |
| trimethyethene | **−15.5** | −25.5 | 4.1 | −14.9 |
| tetramethylethene | **−15.8** | −23.3 | 9.4 | −12.0 |

[a] The 6-31++G** basis set was used for MP2 and B3LYP calculations. The G3MP2, MP2, and B3LYP data were taken from the paper by Check and Gilbert.[16]

agreement with the G3MP2 data (MAE is of 1.1 kcal/mol), whereas MP2 calculations provide less accurate results (MAE is 8.9 kcal/mol);[16] B3LYP performs poorly predicting the reaction energies with MAE of 17.2 kcal/mol.[16]

**3.4. Bond Dissociation Energies.** The C−C and C−H bond energies are often required for the discrimination of reaction mechanisms for complex processes such as hydrocarbon cracking and combustion. Despite the expanding availability of the experimental measurements, the database for the accurate bond energetics is still not comprehensive. Since MTB/2 provides quite accurate values of $\Delta H^0_{f,298}$ for molecules and radicals, it can be used to estimate the enthalpy of bond-breaking and bond-forming reactions. In Table 9, we compare the enthalpies for 14 C−H and 18 C−C bond dissociation reactions. As can be seen, the experimental data are well reproduced by MTB/2. The mean absolute error is 2.45 kcal/mol. It has been shown that B3LYP provides unreliable results for C−C bond energies in branched hydrocarbons,[16] in particular, the bond dissociation energy in $(CH_3)_3C-C(CH_3)_3$ is underestimated by *ca.* 20 kcal/mol.

On average, MTB/2 overestimates the energy of C−H bond (the mean deviation is 1.61 kcal/mol) and underestimates the strength of C−C bond (the mean deviation is −1.63 kcal/mol). Thus, the scheme can be helpful for computer simulations of processes that involve bond-breaking or/and bond-forming reactions in hydrocarbons.

**3.5. Geometries.** Calculated molecular geometries of hydrocarbons and fullerenes agree well with experiment (see Table S2 in the Supporting Information). MTB/2 reproduces accurately bond lengths and bond angles of different types, the corresponding MAEs are 0.013 Å ( 110 comparisons) and 1.3° (38 comparisons). The standard semiempirical methods show similar accuracy: 0.013 Å and 1.4° (AM1) and 0.011 Å and 1.6° (PM3). Such errors seem to be not of great chemical significance, and more accurate estimation of molecular geometries than this is not an important requirement for approximate QC approaches.

**3.6. Comparison of MTB/2 and MTB.** The MTB/2 method shows remarkable improvement over its predecessor MTB.[41] The mean absolute error estimated for $\Delta H^0_{f,298}$ of 146 species of MTB/2 (1.90 kcal/mol) is essentially smaller than that of MTB (2.82 kcal/mol). The new scheme shares with MTB the ability to accurately reproduce heats of formation of alkanes and alkenes. A major improvement has been found in the treatment of polycyclic aromatic molecules, and the MAEs of MTB/2 and MTB are 1.52 and 4.07 kcal/mol. MTB overestimates $\Delta H^0_{f,298}$ of aromatic compounds; for instance, $\Delta H^0_{f,298}$ of chrysene and coronene are predicted to be too high by ca. 10 kcal/mol. In contrast, MTB/2 performs very well for large aromatic molecules (see Table 5); in particular, the errors in $\Delta H^0_{f,298}$ of chrysene and coronene are 2.5 and 1.8 kcal/mol. The MTB/2 $\Delta H^0_{f,298}$ values for radicals are, overall, somewhat better than those from MTB; the corresponding MAEs are 2.47 and 2.84 kcal/mol. Both methods provide satisfactory molecular geometries of hydrocarbons.

***Table 9.*** Bond Dissociation Enthalpies[a]

| reaction | | exp. | MTB/2 |
|---|---|---|---|
| C−H Bond Breaking | | | |
| methane | → $CH_3$+H | 104.9 | 106.3 |
| ethane | → $C_2H_5$+H | 101.0 | 100.0 |
| ethylene | → $C_2H_3$+H | 103.6 | 103.9 |
| propane | → n-$C_3H_7$+H | 100.4 | 99.3 |
| | → i-$C_3H_7$+H | 96.2 | 94.3 |
| propene | → $C_3H_5$+H | 86.6 | 85.5 |
| n-butane | → n-$C_4H_9$+H | 100.2 | 99.4 |
| | → s-$C_4H_9$+H | 99.1 | 94.9 |
| isobutane | → t-$C_4H_9$+H | 93.7 | 90.8 |
| cycloropane | → c-$C_3H_5$+H | 106.3 | 99.8 |
| cyclopropene | → c-$C_3H_3$+H | 90.9 | 92.3 |
| cyclopentadiene | → c-$C_5H_5$+H | 78.0 | 79.7 |
| benzene | → Ph + H | 108.2 | 104.0 |
| toluene | → $PhCH_2$ + H | 89.1 | 86.4 |
| mean error (14 comparisons) | | | 1.54 |
| mean abs. error | | | 1.88 |
| standard deviation | | | 2.39 |
| C−C Bond Breaking | | | |
| ethane | → 2 $CH_3$ | 90.0 | 97.9 |
| propane | → $C_2H_5$+$CH_3$ | 88.9 | 93.4 |
| propene | → $C_2H_3$+$CH_3$ | 94.2 | 100.6 |
| n-butane | → 2$C_2H_5$ | 87.8 | 88.3 |
| | → $C_3H_7$+$CH_3$ | 88.3 | 92.1 |
| isobutane | → i-$C_3H_7$+$CH_3$ | 86.2 | 88.9 |
| butene-1 | → $C_2H_3$+$C_2H_5$ | 92.9 | 94.3 |
| | → Allyl+$CH_3$ | 74.3 | 77.0 |
| methylcyclopropene | → c-$C_3H_3$+ $CH_3$ | 98.9 | 101.5 |
| 2-methylbutane | → i-$C_3H_7$+$C_2H_5$ | 84.7 | 84.0 |
| | → s-$C_4H_9$+$CH_3$ | 88.7 | 89.7 |
| neo-pentane | → t$C_4H_9$+$CH_3$ | 84.7 | 86.0 |
| 2,3-dimethylbutane | →2 i-$C_3H_7$ | 80.7 | 78.5 |
| 2,2,3,3-tetramethylbutane | → 2 t-$C_4H_9$ | 75.2 | 73.4 |
| toluene | → Ph + $CH_3$ | 98.8 | 101.1 |
| ethylbenzene | → $PhCH_2$ + $CH_3$ | 76.9 | 78.1 |
| | → Ph + $C_2H_5$ | 97.6 | 93.9 |
| styrene | → Ph + $C_2H_3$ | 104.5 | 105.1 |
| mean error (18 comparisons) | | | 1.69 |
| mean abs. error | | | 2.15 |
| standard deviation | | | 2.89 |

[a] In kcal/mol.

**3.7. Computational Performance.** The limiting step of semiempirical calculations is the diagonalization of the Hamiltonian matrix. Usually a single-point calculation requires about 20 iterations. It means that 20 diagonalizations are needed for closed-shell systems, while twice as many operations are required for the spin-unrestricted scheme commonly applied to radicals. Moreover, the number of iterations remarkably increases with the size of the model, especially when treating radicals. In such situations, semiempirical calculations become time-consuming. In contrast, a single diagonalization of the Hamiltonian matrix is required within MTB/2, independent of whether a closed- or open-shell system is computed. This makes MTB/2 extremely

efficient for studying large species. For instance, a single-point MTB/2 calculation on a PC of the ground state of $C_{60}$ and $C_{70}$ takes 0.30 and 0.42 s, respectively, whereas AM1 calculations are found to be by a factor of 8 less efficient (for $C_{60}$ and $C_{70}$, they require 2.20 and 3.34 s). As expected, a more considerable gain in computational time is found for open-shell systems. Our MTB/2 and AM1 calculations of the lowest triplet state of $C_{60}$ take 0.30 and 12.40 s, respectively (a factor of 40!).

## 4. Conclusions

We described the MTB/2 approach that provides accurate thermochemical data and structural parameters of hydrocarbons and fullerenes. The mean absolute error of 1.90 kcal/mol is obtained for the $\Delta H^0_{f,298}$ values across various classes of molecules and radicals (146 comparisons). The proposed scheme gives more accurate estimates for reaction enthalpies than MP2, B3LYP, and the standard semiempirical methods. Molecular geometries are well reproduced by MTB/2; the MAEs of 0.013 Å and 1.3° are found for bond lengths and bond angles, respectively.

As the MTB/2 method is based on the noniterative orthogonal scheme, it is computationally very efficient and allows for fast screening of $\Delta H^0_{f,298}$ and reaction enthalpies for large hydrocarbons and fullerenes. Because of its performance, the scheme appears very promising to study the reaction dynamics of hydrocarbons.

In MTB/2, we employ simple functions to construct the effective Hamiltonian matrix (eqs 4 and 5) and short-range interatomic potential (eq 6). The performance of the model is likely to be improved by using more flexible functions to approximate the distance dependence of these terms. Obviously, more accurate thermochemical data can be obtained for a certain class of hydrocarbons (at the cost of less accurate estimates for other compounds) by the corresponding tuning of the MTB/2 parameters.

**Supporting Information Available:** G3, B3LYP, AM1, PM3, and MTB/2 values of $\Delta H^0_{f,298}$ for 16 medium-sized hydrocarbons (Table S1), comparison of the MTB/2 molecular geometries of hydrocarbons with experimental data (Table S2), and a computer program for MTB/2 calculations on PC under the MS Windows environment along with a brief instruction and input examples. This material is available free of charge via the Internet at http://pubs.acs.org.

### References


(1) *NIST Chemistry WebBook*; NIST Standard Reference Database, No. 69; Mallard, W. G., Linstrom, P. J., Eds.; National Institute of Standards and Technology: Gaithersburg, MD.http://webbook.nist.gov (accessed Apr 12, 2008).

(2) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*, 2nd ed.; Chapman and Hall: New York, 1986.

(3) Cohen, N.; Benson, S. W. *Chem. Rev.* **1993**, *93*, 2419.

(4) *Computational Thermochemistry*; Irikura, K. K., Frurip, D. J., Eds.; ACS Symposium Series No. 677, Washington, DC, 1998.

(5) Bond, D. *J. Org. Chem.* **2007**, *72*, 5555.

(6) *Quantum-Mechanical Prediction of Thermochemical Data*; Understanding Chemical Reactivity Vol. 22, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, 2001.

(7) Curtiss, L. A.; Raghavachari, K. *Theor. Chem. Acc.* **2002**, *108*, 61.

(8) Raghavachari, K.; Curtiss, L. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dykstra, C., Frenking, G., Kim, K., Scuseria, G., Eds.; Elsevier: 2005.

(9) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105.

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

(11) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. *J. Phys. Chem. A* **2000**, *104*, 5850.

(12) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(13) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2005**, *123*, 124107.

(14) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.

(15) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.

(16) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.

(17) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(18) Thiel, W. *Adv. Chem. Phys.* **1996**, *93*, 703.

(19) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(20) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275.

(21) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159.

(22) Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *4*, 297.

(23) Goringe, C. M.; Bowler, D. R.; Hernandez, E. *Rep. Prog. Phys.* **1997**, *60*, 1447.

(24) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.

(25) Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397.

(26) Selvam, P.; Tsuboi, H.; Koyama, M.; Kubo, M.; Miyamoto, A. *Catal. Today* **2005**, *100*, 11.

(27) Porezag, D.; Frauenheim, T.; Kohler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.

(28) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

(29) Krüger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 114110.

(30) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006**, *110*, 13551.

(31) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.

(32) Xu, C. H.; Wang, C. Z.; Chan, C. T.; Ho, K. M. *J. Phys. Condens. Matter* **1992**, *4*, 6047.

(33) Wang, C. Z.; Chan, C. T.; Ho, K. M. *Phys. Rev. Lett.* **1991**, *66*, 189.

(34) Wang, Y.; Mak, C. H. *Chem. Phys. Lett.* **1995**, *235*, 37.

(35) Horsfield, A. P.; Godwin, P. D.; Pettifor, D. G.; Sutton, A. P. *Phys. Rev. B* **1996**, *54*, 15773.

(36) Winn, M. D.; Rassinger, M.; Hafner, J. *Phys. Rev. B* **1997**, *55*, 5364.

(37) Pan, B. C. *Phys. Rev. B* **2001**, *64*, 155408.

(38) Zhao, J. J.; Lu, J. P. *Phys. Lett. A* **2003**, *319*, 523.

(39) Wang, C. Z.; Ho, K. M. *J. Comput. Theor. Nanosci.* **2004**, *1*, 3.

(40) Voityuk, A. A. *J. Chem. Theory Comput.* **2006**, *2*, 1038.

(41) Voityuk, A. A. *Chem. Phys. Lett.* **2006**, *433*, 216.

(42) Jasper, A. W.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 210.

(43) Chadi, D. J. *Phys. Rev. B* **1979**, *19*, 2074.

(44) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.

(45) Yu, J.; Sumathi, R.; Green, W. H. *J. Am. Chem. Soc.* **2004**, *126*, 12685.

(46) Sablier, M.; Fujii, T. *Chem. Rev.* **2002**, *102*, 2855.

CT8003222

JCTC Journal of Chemical Theory and Computation

# On the $Pt^+$ and $Rh^+$ Catalytic Activity in the Nitrous Oxide Reduction by Carbon Monoxide

F. Rondinelli, N. Russo,* and M. Toscano

*Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite-Centro d'Eccellenza MIUR, Universita' della Calabria, I-87030 Arcavacata di Rende (CS), Italy*

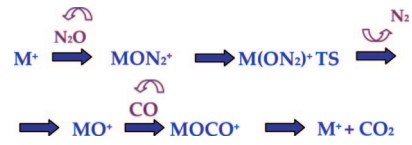
Received May 29, 2008

**Abstract:** Nitrous oxide activation by CO in the presence of platinum and rhodium monocations was elucidated by density functional methods for ground and first excited states. Platinum and rhodium cations fulfill the thermodynamic request for the oxygen-atom transport that allows the catalytic cycle to be completed, but actually, just the first one meaningfully improves the kinetics of the process. For both catalysts, the reaction pathways show the only activation barrier in correspondence of nitrogen release and monoxide cation formation. The kinetic analysis of the potential energy profile, in agreement with ICP/SIFT MS experimental data, indicates that platinum performs more in the reduction, while the whole process is not sufficiently fast in the case of rhodium ionic catalyst.

## Introduction

$N_2O$ is a greenhouse gas that is naturally present in trace amounts in the Earth's atmosphere. The increasing concentration of nitrous oxide drastically influences global warming because it is estimated that each molecule affects Earth's temperature about 300 times more than carbon dioxide. Just recently, its environmental impact was proved. The lack of $N_2O$ elimination processes in the troposphere gives rise to its spread throughout the stratosphere, where it causes the ozone depletion.[1−3] Moreover, contrary to carbon dioxide emissions, less than 40% of atmospheric nitrous oxide comes from anthropic sources. In recent years, since its environmental dangerousness was proven, $N_2O$ emissions control became the object of study of many research areas, from chemistry to engineering. In addition, the new laws about the acceptable levels of outdoor pollutants boosted the introduction of efficient catalytic methods.[4] Although the catalytic process by which transition and main group metals catalyze the conversion of $N_2O$ has not yet been completely investigated, it is, however, well-known that transition and main group metal catalysts do perform considerably well and give rise to versatile reaction events under conditions of homogeneous and heterogeneous catalysis.[5−7] As part of our systematic study[8,9] about the efficiency of experimentally analyzed[6,10] metal monocations, the $N_2O$ activation mechanism by carbon monoxide in the presence of $Rh^+$ and $Pt^+$ was elucidated by DFT approach. The stepwise process proposed by Kappes and Staley[11] is well-known, but just recently, the case of the homogeneous catalysis was considered as a possible reactive event for nitrous oxide conversion. According to mass spectrometry data,[6] rhodium and platinum monocations can both act as oxygen carriers in the mechanism indicated in the Scheme 1, but $Rh^+$ does not kinetically improve the process.

***Scheme 1.*** First and Second Part of $N_2O$ Activation Catalytic Cycle (M = Rh, Pt)




Previous studies on a similar subject,[8,9] have pointed out the influence that in such a type of processes can have the two state reactivity (TSR) phenomenon[12] that consists of the occurrence of crossovers between low- and high-spin energetic profiles. In fact, most of the experimental determinations are based on the catalytic behavior of metal cations only in their ground state. However, often, the participation of metal ions excited states appeared to be of fundamental

* To whom correspondence should be addressed. Fax: +39-0984-493390. E-mail: nrusso@unical.it.

10.1021/ct800199b CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/10/2008

importance to explain the differences in the performance that experiment attributes to the various metal cations and, in some cases, has contributed to better rationalization of some uncertain results because of the difficulty of the experimental measures.[9] Since the TSR concept originated from theoretical studies of a number of gas-phase reactions of small transition metal cations like $Rh^+$ and $Pt^+$, we have made it mandatory to take it into account in this investigation.

## Computational Strategy

All calculations were carried out using the Gaussian 03 program package.[13] The hybrid B3LYP functional[14,15] was adopted in connection with Stuttgart RSC ECP for metals,[16] and 6−311+G(d)[17−19] was used for carbon, nitrogen, and oxygen atoms. All optimizations, carried out without any constraints, were followed by the vibrational analysis at the same theoretical approach to identify the stationary points located on the potential energy surfaces. Zero-point energy correction was added to all the absolute energies. Intrinsic reaction coordinate (IRC) calculation[20] was performed to verify the correction of all transition states. In addition, the stability of DFT density function was tested by the Stable calculation method,[21,22] while th eALTER keyword[13] ensured that the lowest-energy electronic states were found.

## Results and Discussion

As previously mentioned, because of the presence of metal cations as catalysts, particular attention was devoted to the possible occurrence of a two-state reactivity phenomenon.[23−26] Therefore, the potential energy surfaces were traced for rhodium and platinum monocations ground and first excited states. The reliability of the stationary points energetic scale is strictly dependent on the exact gap between the two electronic states of the catalyst. Stuttgart RSC pseudopotential correctly reproduces the stability order between $Rh^+$ and $Pt^+$ most stable excited states.

**(a). $N_2O$ Deoxygenation Catalyzed by Rhodium Monocation.** In agreement with experimental data,[27] the $^3F$ rhodium monocation lowest-energy state is followed by $^1D$, and both correspond to the $4d^8$ configuration. According to our calculations, $^1D$ $Rh^+$ resulted higher than $^3F$ by 52.8 kJ/mol, versus an experimental value of 78.2 kJ/mol.[27] Although in the case of the rhodium electronic states, the theoretical value of the gap is quite different than the experimental one,[27] the reproduction of the spin states' correct order by an adequate basis set is a result rarely obtained by DFT methods.[28]

B3LYP/Stuttgart results indicate that $Pt^+$ $^2D$ $(5d^9)$/$^4F$ $(6s^1 5d^8)$ gap corresponds to 77.0 kJ/mol, in good agreement with the estimated value of 73.0 kJ/mol.[29]

Experimental evidence indicates that both $Rh^+$ and $Pt^+$ can act as oxygen carriers, where their oxygen affinities (OA) are found to be between the $N_2$ and CO affinities.[6] Theoretical values of OA($N_2$) and OA(CO) of 167.9 and 528.8 kJ/mol, respectively, were computed in a previous work[8] by B3LYP/6−311+G(d) theoretical approach and define a range width of 360.8 kJ/mol in which $Pt^+$ and $Rh^+$ oxygen affinities were estimated to be 254.1 and 227.7 kJ/mol,

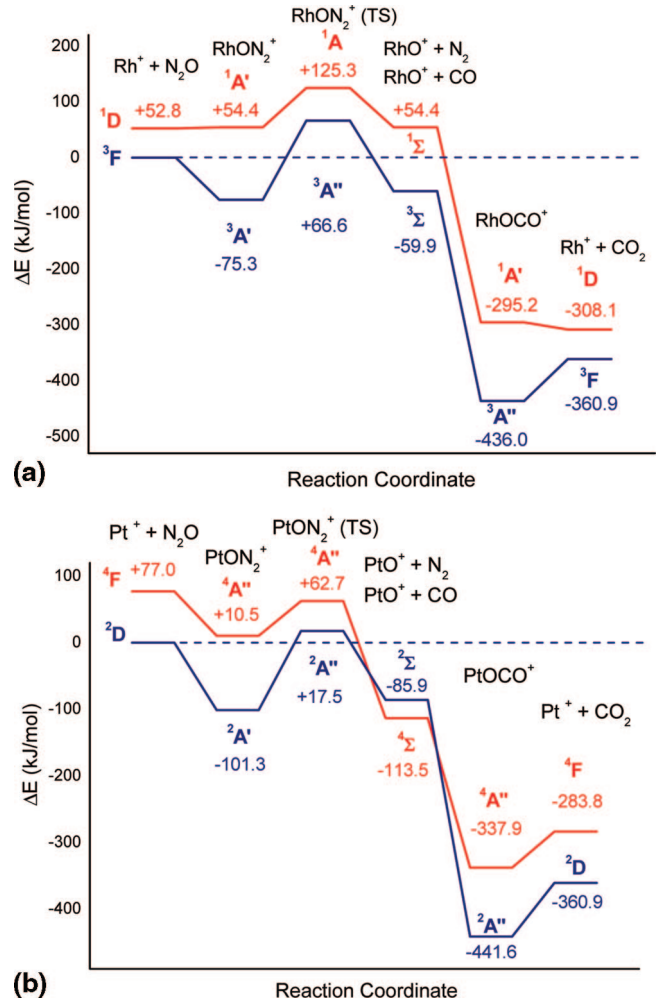



***Figure 1.*** Nitrous oxide activation by carbon monoxide catalyzed by (a) $Rh^+$ and (b) $Pt^+$.

respectively. As it is often found for DFT calculations, these values are underestimated compared to the experimental data of 322.3 and 291.3 ± 5.9 kJ/mol, correspondingly.[30]

Thus, the different catalytic activities of the platinum and rhodium monocations, both thermodynamically suitable to the oxygen transport mechanism, evidently have a kinetic origin as the potential energy surfaces (PES) highlight. In Figure 1, $N_2O$ activation PES for both catalysts is reported. In the case of platinum and rhodium, first and second part of the whole process are characterized by the same steps. The interaction between the metal cation and nitrous oxide allows the formation of the oxygen side-bonded adduct $MON_2^+$ that, through a transition state, evolves into the $MO^+$ species releasing nitrogen. Subsequently, after the introduction of carbon monoxide in the reaction environment, the monoxide cation binds CO by the carbon side, giving rise to $MOCO^+$ complex that releases carbon dioxide regenerating the catalyst. As in our previous works on the same deoxygenation catalytic cycle,[8,9] we verified the coordination compounds to $N_2O$ and CO to be the most stable species. In particular, we found that the metal cations always prefer to bind nitrous oxide by the oxygen side and that $MO^+$ species gives rise to an extremely stable compound through the coordination between the cation's oxygen and CO's carbon.

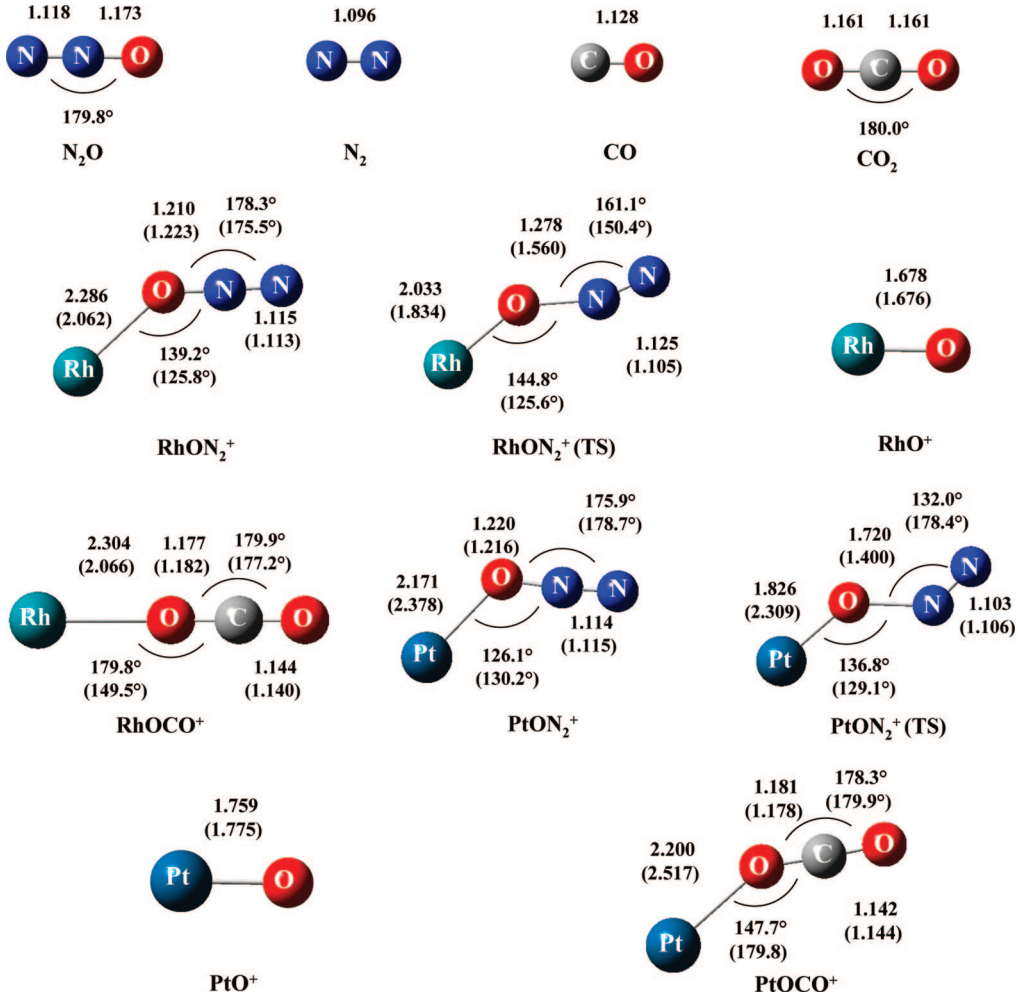



***Figure 2.*** Optimized geometries of rhodium and platinum stationary points in their ground and first excited states. Bond distances are given in angstrom, and angles are given in degrees. Values in parentheses are referred to singlet spin state for rhodium and quartet multiplicity for platinum.

In the case of rhodium, no crossover characterizes triplet and singlet reaction channels because compounds at triplet multiplicity show always a higher stability. Therefore, the process takes place following the ground-state potential energy surface.

Except for the only transition state of the reduction process, all ground-state stationary points lie below the triplet reference reactants. Unlike the first step on singlet PES, the formation of the $^3$A′ $RhON_2^+$ complex is a stage exothermic by 75.3 kJ/mol. The only energetic barrier to overcome corresponds to 66.6 kJ/mol and allows the formation of the monoxide cation. It is worthwhile that in a gas-phase experiment, the intermediates do not lose all of their energy. Thus, the relevant barrier is not that of the bottom of the well, but rather that from the reactants.

The released $N_2$, inert toward the species involved in the process, does not influence the course of reaction. In the saddle point $RhON_2^+$(TS), the normal mode relating to the imaginary frequency of 857.6i cm$^{-1}$ is associated to the O−N stretching corresponding to the breaking of bond. After the introduction of CO, the formation of the $^3\Sigma$ rhodium monoxide cation, stage exoergonic by 59.9 kJ/mol, is followed by the appearance of the $^3$A″ $RhOCO^+$ adduct. The species, product of the $RhO^+$ coordination to CO by its carbon-side, is extremely stable with respect to the reference reactants, feature that we always have found for $MOCO^+$ compounds (M = Fe, Mn, Se, Ge) in our previous studies.[8,9] In the rhodium coordination compound, both O−C distances of 1.14 and 1.18 Å (see Figure 2 for geometrical details) are very close to free carbon dioxide bond lengths, allowing the release of the catalyst in a barrierless step.

Among the first excited-state species, it is worthwhile that $^1$A′ $RhON_2^+$ complex is almost isoenergetic with the reactants including the metal at the same multiplicity. The adduct gives rise to rhodium monoxide cation by overcoming a barrier of 70.9 kJ/mol. $^1$A $RhON_2^+$(TS) imaginary frequency equal to 716.8i cm$^{-1}$ corresponds to O−N bond scission. Carbon monoxide coordination to $^1\Sigma$ $RhO^+$ leads to the formation of $^1$A′ $RhOCO^+$ compound, 12.9 kJ/mol less stable than the products located on the path at the same multiplicity.

As any interaction characterizes ground and first excited states potential energy surfaces, the reaction mediated by rhodium cation is a single-state reactivity process and takes place along the triplet path. The activation barrier of 66.6 kJ/mol prevents the formation of an adequate quantity of rhodium monoxide cation to complete the catalytic cycle. This is the origin of the impossibility to experimentally detect

$RhO^+$ in the reaction environment.[6] Therefore, this stage of the process compromises $Rh^+$ catalytic activity. The remarkable reaction rate improvement induced by platinum was quantified by Bohme and co-workers through the valuation of a first step kinetic constant $10^3$ higher than rhodium one.[6]

**(b). $N_2O$ Deoxygenation Catalyzed by Platinum Monocation.** In agreement with this data, the nitrous oxide reaction path in the presence of platinum, illustrated in Figure 1b, shows a very low activation barrier. Doublet and quartet reaction channels show two adjacent spin inversions before and after $PtO^+$ appearance, subsequently the only transition state, so that strictly we cannot define the reaction as a two-state reactivity process.[12] Along the doublet reaction path, $N_2O$ coordination to the metal catalyst gives rise to the $^2A'$ $PtON_2^+$ species, 101.3 kJ/mol lower than the ground-state reactants. In its excited state, the same compound is 10.5 kJ/mol above the reactants asymptote. The release of nitrogen is preceded by the $^2A''$ transition state for which the vibrational analysis gave an imaginary frequency of 401.3 $cm^{-1}$. The $^4\Sigma$ platinum monoxide cation, located at 113.5 kJ/mol below the reactants asymptote is about 27.6 kJ/mol more stable than the doublet species so that, after the $^2A''$ transition state, the reaction moves temporarily along the excited energetic profile rather than follow the doublet surface. The second crossing allows the catalytic cycle to regain the ground-state reaction channel being the $^2A''$ $PtOCO^+$ strongly favored (of about 103.7 kJ/mol) with respect to the corresponding quartet species.

In the case of quite exothermic deoxygenation processes concerning the nitrous oxide activation mechanism, the spin conservation resulting from the interaction of potential energy surfaces was already mentioned by Schwarz.[31]

The detachment of carbon dioxide from $^2A''$ $PtOCO^+$ compound is a step that is exothermic by 360.9 kJ/mol. The drastic decrease of the $N_2O$ reduction activation barrier, from 199.2 kJ/mol[8] in the absence of catalyst to 17.5 kJ/mol because of the interaction with $Pt^+$, indicates in agreement with experimental data[6] the excellent performance of platinum monocation.

A glance to the pathways shown in Figure 1 reveals that, as far as the oxide formation is concerned, the reaction profiles for both examined cations are quite similar to that obtained by Bohme and co-workers[10] for the reaction mediated by iron catalyst. However, the second part of the catalytic cycle appears different because of the absence of the $OCMO^+$ species and of the next transition state that should evolve into the $MOCO^+$ most stable adduct. In our previous study concerning the iron catalytic activity in the reduction of $N_2O$ by CO,[8] in which we took account some aspects of the problem not investigated previously, we ascertained that the lack of the $OCMO^+$ stationary point along the reaction's PES does not influence neither kinetics nor thermodynamic of the process since the energetic profile after the oxide formation lies entirely below the reactants asymptote.

## Conclusions

On the basis of an experimental screening about the performances of some atomic cations carried out by Bohme et al.,[6] the oxygen transport activation of nitrous oxide by carbon monoxide mediated by $Pt^+$ and $Rh^+$ was analyzed by B3LYP/Stuttgart RSC ECP/6−311+G(d) on the basis of the mechanism suggested by Kappes and Staley.[11] The evaluation of the thermodynamic condition necessary to the achievement of the catalytic cycle, common to both metal cations, was followed by the analysis of the kinetic features of the process. Ground and first excited-state reaction channels were traced to consider possible spin inversions in the activation process. The reliability of our data is demonstrated by the exact ordering of catalysts' spin states. The stepwise mechanism in the case of rhodium catalyst entirely takes place along the triplet ground-state PES. The lowering of the activation barrier with respect to the uncatalyzed process is not enough significant to give rise to a sufficient $RhO^+$ production so that, as indicated by experimental evidence, the second and last part of the catalytic cycle cannot easily take place.

In the case of platinum monocation, except for a double spin inversion in correspondence of $PtO^+$ formation, the reaction proceeds along the doublet ground-state channel. The global process is therefore characterized by reactants and products in the same multiplicity. According to our analysis and in agreement with experimental evidence, $Pt^+$ performs in the catalysis of $N_2O$ reduction as the activation barrier to overcome in the first part of the process is equal to 17.5 kJ/mol.

**Acknowledgment.** We gratefully acknowledge the Dipartimento di Chimica, Università della Calabria, for financial aid.

## References

(1) Gonçalves, F.; Figueiredo, J. L. *Appl. Catal., B* **2004**, *50*, 271.

(2) Armor, J. A. *Appl. Catal., B* **1992**, *1*, 221.

(3) Dandekar, A.; Vannice, M. A. *Appl. Catal., B* **1999**, *22*, 179.

(4) Tomita, A. *Fuel Process. Technol.* **2001**, *71*, 53.

(5) Sun, K.; Xia, H.; Hensen, E.; van Santen, R.; Li, C. *J. Catal.* **2006**, *238*, 186.

(6) Blagojevic, V.; Orlova, G.; Böhme, D. K. *J. Am. Chem. Soc.* **2005**, *127*, 3545.

(7) Burch, R.; Daniells, S. T.; Breen, J. P.; Hu, P. *J. Catal.* **2004**, *224*, 252.

(8) Rondinelli, F.; Russo, N.; Toscano, M. *Inorg. Chem.* **2007**, *46*, 7489.

(9) Chiodo, S.; Rondinelli, F.; Russo, N.; Toscano, M. *J. Chem. Theory Comput.* **2008**, *4*, 316.

(10) (a) Lavrov, V. V.; Blagojevic, V.; Kojanagi, G. K.; Orlova, G.; Bohme, D. K. *J. Phys. Chem. A* **2004**, *108*, 5610. (b) Böhme, D. K.; Schwarz, H. *Angew. Chem., Int. Ed.* **2005**, *44*, 2336.

(11) Kappes, M. M.; Staley, R. H. *J. Am. Chem. Soc.* **1981**, *103*, 1286.

(12) Schröder, D.; Shaik, S.; Schwarz, H. *Acc. Chem. Res.* **2000**, *33*, 139.

(13) 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.; Ispida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strani, 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, V.; 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 program*, version B04; Gaussian, Inc, Wallingford, CT, 2004.

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

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

(16) Basis sets were obtained from the Extensible Computational Chemistry Environment Basis Set Database, version 02/02/06, as developed and distributed by the Molecular Science Computing Facility, Environmental and Molecular Sciences Laboratory which is part of the Pacific Northwest Laboratory, P.O. Box 999, Richland, WA 99352, and funded by the U.S. Department of Energy. The Pacific Northwest Laboratory is a multi-program laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06−76RLO 1830.

(17) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(18) Clark, T.; Chandrasekhar, J.; Schleyer, P. V. R. *J. Comput. Chem.* **1983**, *4*, 294.

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

(20) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(21) Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1977**, 66.

(22) Bauernschmitt, R.; Ahlrichs, R. *J. Chem. Phys.* **1996**, *104*, 9047.

(23) Armentrout, P. B.; Beauchamp, J. L. *Acc. Chem. Res.* **1989**, *2*, 315.

(24) Armentrout, P. B. *Science* **1991**, *251*, 175.

(25) Fiedler, A.; Schröder, D.; Shaik, S.; Schwarz, H. *J. Am. Chem. Soc.* **1994**, *116*, 10734.

(26) Shaik, S.; Danovich, D.; Fiedler, A.; Schröder, D.; Schwarz, H. *Helv. Chim. Acta* **1995**, *78*, 1393.

(27) Mandich, M. L.; Halle, L. F.; Beauchamp, J. L. *J. Am. Chem. Soc.* **1984**, *106*, 4403.

(28) Chiodo, S.; Russo, N.; Sicilia, E. *J. Comput. Chem.* **2004**, *26*, 175.

(29) Moore, C. E. *At. Energy Levels* **1971**, *35*.

(30) Lavrov, V. V.; Blagojevic, V.; Gregory, K. K.; Orlova, G.; Böhme, D. K. *J. Phys. Chem.* **2004**, *108*, 5610.

(31) Schwarz, H. *Int. J. Mass Spectrom.* **2004**, *237*, 75.

CT800199B

JCTC Journal of Chemical Theory and Computation

# Toward a Coarse-Grained Protein Model Coupled with a Coarse-Grained Solvent Model: Solvation Free Energies of Amino Acid Side Chains

Wei Han,[†] Cheuk-Kin Wan,[†] and Yun-Dong Wu*[,†,‡]

*Department of Chemistry, The Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong, China, and National Laboratory of Molecular Sciences, College of Chemistry, Peking University, Beijing, China*

Received May 23, 2008

**Abstract:** Recently, we reported that molecular dynamics (MD) simulations using a coarse-grained (CG) peptide model coupled with a CG water model are able to reproduce many of the structural and thermodynamic features of short peptides with nonpolar side chains at $10^3$ times the normal speed (*JCTC*, **2007**, *3*, 2146−2161). To further develop a CG protein model for MD simulations, we systematically parametrized the side chains of all 20 naturally occurring amino acids. We developed the parameters by fitting the dihedral potentials of 13 small molecules, the densities and self-solvation free energies of liquids of eight organic molecules, and the hydration free energies of 35 small organic molecules. In a set of 11 classes of compounds (105 in total) including alkanes, alcohols, ethers, ketones/aldehydes, amines, amides, aromatics, carboxylic acids, sulfides/thiols, alkyl ammoniums, and carboxylate ions, the average error in the calculated hydration free energies compared with experimental results is about 1.4 kJ/mol. The average error in the calculated transfer free energies of the 19 side-chain analogues of amino acids from cyclohexane to water is about 2.2 kJ/mol. These results are comparable to the results of all-atom models.

## 1. Introduction

Coarse-grained (CG) force fields have become promising tools for studies of protein folding and protein−protein interactions.[1−12] Unlike the widely used all-atom force field, CG force fields represent a group of atoms with a single particle. The simplicity of the model and smoother potential energy surfaces make CG force fields fast enough to simulate protein folding at biologically relevant time scales, which is currently difficult for all-atom force field models. There have been many applications of CG force fields to protein simulations at various levels. CG models, which explicitly represent each heavy atom, have been used to study the folding of small peptides with about 20 amino acids, such as polyalanine and Trp-cage.[1,2] CG models with coarser resolution have been used to study the formation of helix bundles and the aggregation of peptides.[3−5] A residue-based CG model is even feasible for the study of the stability and dynamics of a viral capsid.[6] Moreover, Voth and co-workers have recently developed a systematical way to derive interaction parameters between different scales by matching CG interaction force with the simulated force in all-atom simulation.[13] This method allows the CG model to study a multiscale system with solvent or membrane environments as coarse-grained model but with proteins represented in more detail.[14−16]

The quality of CG simulations of proteins relies on the interparticle interaction parameters, particularly for hydrogen bonds (H-bonds) and hydrophobic interactions. These interactions are known to depend strongly on the local environment, such as the solvation level, and to have many-body characters.[10,17] They are pairwise additive in many CG force fields. So far, there has not been much effort in parametrize CG force fields with solvation properties except for a recent CG force field for the simulations of biomolecules in lipid/

* Corresponding author e-mail: chydwu@ust.hk.
[†] The Hong Kong University of Science & Technology.
[‡] Peking University.

10.1021/ct800184c CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

water environment, namely the MARTINI force field, which was developed by Marrink and co-workers.[18−20] The way of parametrization in this model has been recognized to be critical in reproducing the thermodynamic properties of systems in a comparative study with all-atom simulation.[21] Their CG force field has been successfully used to study the formation and fusion of micelles and the phase transitions of membranes.[22−25] By including a residue-based protein model, the MARTINI protein model can reproduce the transfer free energies of 16 amino acids side-chain analogues from water to the center of lipid bilayer with an average error of about 8 kJ/mol with respect to all-atom simulations.[19] This model allows the CG model to study interactions between proteins and lipid membrane and dynamics and thermodynamics of proteins in membranes beyond microsecond time scale.[26−32]

Biological processes such as protein folding, protein−protein recognition, and aggregation of transmembrane helices are all associated with the transfer of amino acids from a polar environment such as water to a nonpolar environment.[33−36] Thus, solvation properties of amino acids, particularly the solvation changes in different media, are very important in controlling these processes.[37−41] Indeed, experimental solvation properties of various organic compounds have been used to parametrize several all-atom force fields such as OPLS and GROMOS.[42−47] Recently, the solvation free energies of side-chain analogues of amino acids were studied by OPLS and GROMOS all-atom force fields simulations.[48−50] These simulations reproduced transfer free energies of side-chain analogues of amino acids from cyclohexane to water determined experimentally with reasonable accuracy (2.4−4.2 kcal/mol).

We recently reported our initial effort in developing a CG protein force field model[51] that couples with the CG solvent model developed by Marrink et al.[18] Our model represents amino acid backbone by four particles to include more detail. We developed parameters for peptide backbones, aliphatic side chains, and protein-CG water interactions based on solvation free energies of small molecules. Initial application of the model to the simulation of Ac-$(Ala)_6$-X-$(Ala)_7$-NHMe, X = Ala, Gly, Val, and Leu, indicated the following: (1) this model is 100−1000 times faster than all-atom models; (2) various secondary structures such as helices, turns, and sheet structures can be generated; (3) the helix-coil transition free energies of these peptides are well reproduced; and (4) the structural information from the simulations allows the analysis of the difference in the helical propensities of different amino acids. In this paper, we report on the extension of the parametrization to all 20 naturally occurring amino acid residues. The parametrization is mainly based on solvation properties and can be divided into three parts: (1) all geometric and dihedral parameters are optimized based on conformational energies of organic compounds by quantum mechanics calculations;[46,52] (2) nonbonded interactions are then parametrized by fitting the densities and self-solvation free energies of pure organic liquids and the solvation free energies of organic compounds in water; and (3) the transfer free energy of side-chain analogues of various amino acids from cyclohexane and water were also computed and used to benchmark the performance of our model. We show that the parameters are transferable and the accuracy of the model in determining the hydration free energies of 105 small molecules is comparable to the accuracy of all-atom models.

**2.1. The Coarse-Grained (CG) Protein Model.** The CG protein model has been described in detail in a previous work.[51] Briefly, our protein model is essentially a united-atom model with each CG particle representing a single heavy atom together with the hydrogen atoms attached to it. The potential energy of the system includes bonded and nonbonded terms as shown below:

$$V_{total} = V_{bonded} + V_{non-bonded} \tag{1}$$

The bonded terms ($V_{bonded}$) include quadratic potentials for angle bending ($V_{angle}$) or to conserve $sp^2$ planarity and $sp^3$ configuration ($V_{improper}$) and a combination of two types of potentials ($V_{torsion}$ and $V_{14,ij}$) to describe dihedral angles:

$$V_{Angle} = K_{Angle}(\theta - \theta_0)^2/2 \tag{2}$$

$$V_{Improper} = K_{Improper}(\xi - \xi_0)^2/2 \tag{3}$$

$$V_{torsion} = K_{torsion}(1 + \cos(n\varphi - \varphi_0)) \tag{4}$$

$$V_{14,ij} = \sum_{1-4relationship} 4\varepsilon_{14,ij}\left(\frac{\delta_{14,ij}^{12}}{r^{12}} - \frac{\delta_{14,ij}^{6}}{r^{6}}\right) \tag{5}$$

Both $K_{angle}$ and $K_{improper}$ take the same values, 300 kJ/mol/rad$^2$, as in our previous work. In addition, all bonds are constrained at their equilibrium length ($r_0$) by the LINCS algorithm.[53] Finally, all the nonbonded terms are expressed as Lennard-Jones (LJ) potentials:

$$V_{non-bonded} = \sum_{i<j} 4\varepsilon_{ij}\left(\frac{\delta_{ij}^{12}}{r^{12}} - \frac{\delta_{ij}^{6}}{r^{6}}\right) \tag{6}$$

**2.2. Simulation Setup.** The simulations were performed with the GROMACS 3.3.1 package.[54] For all simulations, the van der Waals (vdW) interactions have a cutoff of 1.2 nm, and they were smoothed to zero from 0.9 to 1.2 nm. The temperature and pressure are controlled by a thermostat and a pressure bath, with coupling constants of 0.1 and 0.5 ps, respectively.[55] The time interval to integrate the Newton equations is 10 fs, and the neighboring list is updated every 10 steps. In the preparation stage, the whole system is subjected to 5000 steps of steep descent optimization and then to a 1200 ps of pre-equilibrium at 300 K and 1 atm. The generated coordinates are used for free energy calculations.

**2.3. Solvation Free Energy Calculations.** The solvation free energy, $\Delta G_{sov}$, is defined as the free energy difference between the state where the solute is immersed in the solvent and the state where the solute is isolated from the solvent. The free energy perturbation method calculates the solvation free energy by introducing a coupling parameter ($\lambda$) with interaction potentials between the solute and the solvent. As $\lambda$ gradually varies from zero to unity, the interactions are gradually turned on. During this process, $\Delta G_{sov}$ can be calculated as[56]

$$\Delta G_{sov} = G_1 - G_0 = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{7}$$

***Table 1.*** Summary of Uncertainties of Calculated Solvation Free Energies of Various Solution Systems

| solution systems | CG particle number | total time of perturbation (ns) | uncertainties (kJ/mol) |
|---|---|---|---|
| pure liquids[a] | 1000−2000 | 128 | <1 |
| water solutions[b] | 350 | 300 | 0.2−1.5 |
| cyclohexane solutions[c] | 1250 | 128 | <1 |

[a] Pure liquids of cyclohexane, n-pentane, isopentane, 2,3-dimethyl-2-butene, diethyl ether, triethylamine, benzene, and dimethylsulfide were used to derive their self-solvation free energies. [b] The solutes include 105 organic molecules. [c] The solutes include 19 side-chain analogues of amino acids.

where $U(\lambda)$ is the total energy when the system is in the intermediate state, $\lambda$. For each perturbation calculation, there are 32 intermediates ($\lambda$=0.0, 0.02, 0.06, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.44, 0.47, 0.5, 0.53, 0.56, 0.59, 0.62, 0.65, 0.68, 0.71, 0.73, 0.75, 0.78, 0.81, 0.84, 0.87, 0.9, 0.92, 0.94, 0.96, 0.98, 1.0). In addition, a soft-core Lennard-Jones potential is applied to avoid the singularity problem when $\lambda$ is close to unity or zero.[57]

We carried out free energy calculations for several solution systems including pure liquids of eight organic compounds, aqueous solutions of 105 organic compounds, and cyclohexane solutions of 19 side-chain analogues of amino acids. The detailed information is given in Table 1. The uncertainties of the calculations are estimated by performing three calculations under the same conditions with different initial atomic velocities. The error is larger for water than for nonpolar solvent. It also increases with the number of internal degrees of freedom of the solutes. All the reported solvation free energies were averaged results. As shown in Table 1, the uncertainties of our calculations are less than 1 kJ/mol in most cases, except for water (0.2−1.5 kJ/mol). The calculated uncertainties are comparable to those from similar solvation free energy calculations using all-atom force fields reported by Chang et al. (0.4−1.2 kJ/mol)[50] and MacCallum and Tieleman (less than 2.5 kJ/mol).[48]

## 3. Results and Discussion

There are 24 types of protein particles in our current model used to represent various kinds of amino acids (Table 2). Each of them represents a single heavy atom or a single heavy atom with its attached hydrogen atoms. To couple our model with the CG solvent model developed by Marrink et al.,[18] we also included the CG water particle type, which includes four water molecules.

Therefore, the optimization of the parameters of these particle types was divided into two steps. First, the interaction parameters among protein particle types were optimized. Then the interactions between the CG water particles and protein particles were parametrized. The interaction parameters between the CG water particles were taken directly from the reported results by Marrink et al.[18]

The parametrization of the interactions among the protein particles was further divided into three steps: (1) the equilibrium bond length, $r_0$, and angle value, $\theta_0$, were obtained from the optimized geometries of organic molecules with quantum calculations; (2) the dihedral terms ($\epsilon_{14}$, $\delta_{14}$, $K_{\text{torsion}}$, $n$, and $\phi_0$) were then optimized by fitting the conformational energies of the organic molecules with quantum calculations; and (3) the nonbonded terms ($\epsilon$ and $\delta$) were optimized by fitting the liquid densities and self-solvation free energies of several organic compounds.

***Table 2.*** Summary of Particle Types

| type | description |
|---|---|
| **$CH_4$** | $sp^3$ carbon in methane |
| **-$CH_3$** | $sp^3$ carbon with three H |
| **-$CH_2$-** | $sp^3$ carbon with two H |
| **>CH-** | $sp^3$ carbon with one H |
| **-CH=** | aromatic CH group |
| **>C=O** | carbonyl carbon |
| **-$COO^-$** | carboxylate carbon |
| **-$CH_x$-P** | $sp^3$ carbon directly connected to polar groups (P) |
| **>N-** | $sp^3$ nitrogen without H |
| **>NH** | $sp^3$ nitrogen with one H |
| **-$NH_2$** | $sp^3$ nitrogen with two H |
| -CO-**N**< | N-disubstituted amide |
| -CO-**NH**- | N-monosubstituted amide or heterocycle NH |
| -CO-**$NH_2$** | amide $NH_2$ |
| **=NH** | NH doubly bonded to C |
| **=N-** | heterocycle N |
| **-$NH_3^+$** | ammonium $NH_3^+$ |
| **-OH** | hydroxyl group |
| -CO-**OH** | OH group of carboxylic acid |
| **-O-** | ether oxygen |
| >C=**O** | carbonyl oxygen |
| -C**OO**$^-$ | carboxylate oxygen |
| **-SH** | thiol SH |
| **-S-** | sulfide S |
| **W** | CG water |

**3.1. Optimization of Geometric Parameters.** All the geometric parameters including bond lengths, $r_0$, and angle values, $\theta_0$, were obtained by the optimization of 15 small organic molecules via quantum mechanics calculations with the GAUSSIAN 03 program package.[58] All the calculations were performed at the RHF/6−31G* level. These small molecules and the derived parameters are given in Figure S1 in the Supporting Information (SI).

**3.2. Optimization of Dihedral Parameters.** Thirteen organic molecules were used to derive the dihedral terms including $\epsilon_{14}$, $\delta_{14}$, $K_{\text{torsion}}$, $n$, and $\phi_0$, while the optimized parameters are given in Table 3. The simulated conformational energies and the calculated data by Jorgensen et al. (RHF/6−31G*)[46] and by Halgren[52] (MP4SDQ/TZP)[59,60] are given in Table 4. The conformers of the small molecules can be divided into two categories: stable minima, such as *gauche* (±60°) and *trans* (±180°) conformers of n-butane, and transition states, such as *cis* (0°) and *skew* (±120°) conformers of n-butane. The average energy deviation of the stable minima in our calculations from those by Jorgensen et al. was about 1.4 kJ/mol, mainly resulting from the large deviations of *gauche-trans* relative energies in n-butane, methyl ethyl ether, and propanethiol, which were about 2.0, 3.7, and 1.8 kJ/mol, respectively. However, when compared with the data from the MP4SDQ/TZP calculations, the above three deviations dropped to 0.7, 2.6, and 0.3 kJ/mol, respectively. Also, the deviations for n-butane and methyl ethyl ether from experimental values were 0.6 and 2.9 kJ/mol, respectively.[61,62] Compared with the energies of the transition states from Jorgensen et al., the deviation in our results was about 5.0 kJ/mol. The barriers calculated from our model were systematically lower than those from the

***Table 3.*** Parameters for $\epsilon_{14,ij}$, $\delta_{14,ij}$, $\phi_0$, $K_{torsion}$, and $n$

| pair type | $\epsilon_{14,ij}$ (kJ/mol) | $\delta_{14,ij}$ (nm) |
|---|---|---|
| $C_{sp3}$—$C_{sp3}$[a] | 0.1 | 0.360 |
| -**O**—/—**OH**/-CO-**OH**—$C_{sp3}$ | 0.6 | 0.300 |
| >C=**O**/-C**OO**$^-$—$C_{sp3}$ | 0.6 | 0.280 |
| N[b]—$C_{sp3}$ | 0.1 | 0.345 |
| >**C**=O/-C**OO**$^-$/>**C**<—$C_{sp3}$ | 0.1 | 0.320 |
| -**S**—/—**SH**—$C_{sp3}$ | 0.1 | 0.380 |

| torsion type | $\phi_0$(°) | $K_{tor}$ (kJ/mol) | $N$ |
|---|---|---|---|
| Y[c]-$C_{sp3}$-$C_{sp3}$/$O_{sp3}$/$N_{sp3}$[d]-Y | 0 | 4.9 | 3 |
| Y—S—$C_{sp3}$-Y | 0 | 3.9 | 3 |
| $C_{sp3}$-$C_{sp3}$-$N_{sp2}$-(C=NH) | 0 | 3.0 | 3 |
| $C_{sp3}$-$C_{sp3}$-(CO)-$O^-$/$C_{sp3}$ | 180 | 2.8 | 2 |
| $C_{sp3}$-$C_{sp3}$-(CO)-NH/$NH_2$ | 180 | 0.8 | 2 |
| $C_{sp3}$-$N_{sp2}$-(CO)-$C_{sp3}$ | 180 | 42.0 | 2 |
| $C_{sp3}$-$N_{sp2}$-(C=NH)-$NH_2$ | 180 | 42.0 | 2 |

[a] $C_{sp3}$ indicates -$CH_3$, >$CH_2$, >CH-, and -**CH$_x$**-P. [b] N indicates all types of nitrogen containing groups. [c] Y indicates any type of particles. [d] $O_{sp3}$ and $N_{sp3}$ indicate all types of $sp^3$ oxygen and nitrogen, respectively.

RHF/6−31G* calculations, indicating that it may be easier to change the conformations of the molecules in our model. However, this may have little effect on the equilibrium conformational properties. Alternatively, the RHF/6−31G* calculations may overestimate the barriers. Halgren has pointed out the inaccuracy of RHF/6−31G* in calculating the energies of conformers.[52] For example, the experimental value of the *cis-trans* energy difference of n-butane was actually about 20.4 kJ/mol,[62] lower than that (25.9) from the RHF calculations but closer to ours (17.7). However, the barrier heights of various small molecules still need to be estimated by quantum mechanics calculations at a higher level such as by MP4SDQ/TZP to further refine our dihedral potentials.

**3.3. Optimization of Nonbonded Interactions among Protein Particles.** Our CG model separately treats electrostatic interactions between protein particles such as hydrogen bonds by using other special interaction potentials, as described in our previous work.[51] Thus, our current task is only to obtain van der Waals (vdW) parameters ($\epsilon$ and $\delta$) for interactions between protein particles (eq 6). Adopting the method used in the parametrization of OPLS all-atom force fields,[42−46] we optimized these nonbonded parameters by simulating liquid of organic compounds and fitting their experimental thermodynamic properties.

Specifically, $\delta$ captures the size of a particle and is related to the density of liquid composed of this type of particle. The $\epsilon$ captures the interaction strength between the particles, and it is connected to the solvation free energy of a particle in the solvent. Thus, it is possible to optimize $\epsilon$ and $\delta$ by fitting experimental liquid density and solvation free energies. In principle, all $\epsilon_{ij}$ and $\delta_{ij}$ between any two types of particles could be optimized. In our optimization procedure, we follow a simple approach. We first optimize the parameters between the same protein particle types ($\epsilon_{ii}$ and $\delta_{ii}$) from the liquid densities and the self-solvation free energies of the organic compounds. Then, we derive the parameters for two different particle types ($\epsilon_{ij}$ and $\delta_{ij}$) according to the Lorentz−Berthelot (LB) combination rule:

$$\delta_{ij} = {}^1\!/_2(\delta_{ii} + \delta_{jj}) \tag{8}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \tag{9}$$

For vdW-type interactions, the rules are physically reasonable and mathematically convenient.[63] They have been widely used in all-atom force field models. To reduce the number of parameters to be optimized, the parameters for some protein particle types ($\epsilon_{ii}$ and/or $\delta_{ii}$) were directly used for other protein particle types. The validity of the transfer of parameters was examined by simulating the solvation of various compounds in cyclohexane, which is discussed in Section 3.6.

The interactions for alkyl-type particles including primary (-**CH$_3$**), secondary (>**CH$_2$**), and tertiary (>**CH**-) carbon were first parametrized since alkyl groups are the main parts of various amino acid side chains. Cyclohexane, which has only $CH_2$ groups, was used to derive the parameters for >**CH$_2$**. With the parameters for >**CH$_2$** and applying the LB combination rule, the parameters for -**CH$_3$** were obtained from the simulations of n-pentane, which has both $CH_2$ and $CH_3$ groups. In a similar manner, the parameters for >**CH**- were obtained from the simulations of isopentane. In addition, the parameters for aromatic carbon type (=**CH**-) were obtained by simulating liquid benzene.

For particle types containing O, N, and S, we used aprotic solvents instead of protic solvents to derive their nonbonded parameters. Since we were interested only in the vdW parameters between these particles, this approach avoids the complication of electrostatic interactions such as intermolecular hydrogen bonds, which are present in protic solvents.

We chose diethyl ether for $sp^3$ oxygen (-**O**-). Diethyl ether is the smallest molecule appropriate for our purpose since it has a measurable liquid−gas equilibrium vapor pressure to calculate the self-SFE experimentally. As protic solvents cannot be used, the parameters for -**OH** cannot be directly derived. As an approximation, the $\epsilon_{ii}$ and $\delta_{ii}$ of the -**OH** group were considered the same as those of $sp^3$ oxygens, given that the vdW interaction of -**OH** is mainly from the oxygen.

Carbonyl groups and carboxylic groups that have $sp^2$ carbons (>**C**=O and -**C**OO$^-$) and oxygen (>C=**O** and -C**OO**$^-$) are important parts of peptides. The parameters for the carbonyl group can be derived from simulations of acetone. Nevertheless, the dipole−dipole electrostatic interaction among acetone molecules is considerable in liquid acetone since an acetone molecule has a dipole moment of 2.91 D, which should be avoided in this work. Thus, we obtained the vdW parameters for the $sp^2$ carbon by fitting the experimental data of 2,3-dimethyl-2-butene. For $sp^2$ oxygen, the $\delta_{ii}$ was taken from the crystal structure survey by Chothia and co-workers,[68] which is 0.280 nm, while the $\epsilon_{ii}$ was assumed to be the same as that of an -**O**- group.

In a similar manner, the vdW $\epsilon_{ii}$ and $\delta_{ii}$ of $sp^3$ nitrogen (>**N**-) were obtained through simulations of pure triethylamine (TEA). These vdW parameters were also applied to all other kinds of $sp^3$ and $sp^2$ nitrogens. This assumption may be reasonable since the statistical survey of crystal structures reveals that various kinds of $sp^3$ and $sp^2$ nitrogen groups share the same vdW radii.[68] Finally, the $\epsilon_{ii}$ and $\delta_{ii}$ of

***Table 4.*** Relative Energies (kJ/mol) of Different Conformers of Various Organic Molecules Computed by our CG Model and the Quantum Calculations at the RHF/6-31G*46 and MP4SDQ/TZP[52] Levels

| compound | dihedral | conformer (degree) | coarse grained | RHF/ 6−31G* | MP4SDQ/TZP |
|---|---|---|---|---|---|
| butane | C−C−C−C | 0 | 17.7 | *25.9* | N/A |
| | | 60 | 2.2 | 4.2 | 2.7 |
| | | 120 | 9.9 | 15.3 | N/A |
| | | 180 | 0 | 0 | 0 |
| propanol | C−C−C−O | 0 | 14.7 | 22.6 | N/A |
| | | 60 | 0.4 | 0 | 0.5 |
| | | 120 | 9.3 | 16.4 | N/A |
| | | 180 | 0 | 0 | 0 |
| ethyl methyl ether | C−C−O−C | 0 | 21.1 | 28.6 | N/A |
| | | 60 | 3.3 | 7 | 5.9 |
| | | 180 | 0 | 0 | N/A |
| propyl- ammonium ion | C−C−C−N | 0 | 15.9 | 22.6 | N/A |
| | | 60 | 1.3 | 2.04 | 0.3 |
| | | 120 | 9.8 | 15.9 | N/A |
| | | 180 | 0 | 0 | 0 |
| propylamine | C−C−C−N | 0 | 15.9 | 23.7 | N/A |
| | | 60 | 1.6 | 2.3 | 1.4 |
| | | 120 | 9.9 | 17.1 | N/A |
| | | 180 | 0 | 0 | 0 |
| butanone | C−C−C−O | 0 | 0 | 0 | 0 |
| | | 110 | 5.7 | 7 | 4.1 |
| | | 180 | 12.4 | 12.2 | N/A |
| proanamide | C−C−C−N | 0 | 6.6 | 7.1 | N/A |
| | | 180 | 0 | 0 | N/A |
| butanamide | C−C−C−C(O) | 0 | 17.9 | 24.2 | N/A |
| | | 60 | 1.9 | 1.1 | N/A |
| | | 120 | 10.5 | 12.2 | N/A |
| | | 180 | 0 | 0 | N/A |
| propanoate ion | C−C−C−O | 0 | 0 | 0 | N/A |
| | | 90 | 3.7 | 3 | N/A |
| butanoate ion | C−C−C−C(O) | 0 | 20.7 | 24.2 | N/A |
| | | 60 | 1.8 | 0.3 | N/A |
| | | 120 | 12.5 | 11.3 | N/A |
| | | 180 | 0 | 0 | N/A |
| N-propaneguanidine | C−C−N−C | 0 | 28.6 | 34.8 | N/A |
| | | 90 | 4.0 | 2.8 | N/A |
| | | 120 | 5.7 | 5.4 | N/A |
| | | 180 | 0 | 0 | N/A |
| propanethiol | C−C−C−S | 0 | 19.5 | 26.7 | N/A |
| | | 60 | 2.1 | 3.9 | 1.8 |
| | | 120 | 9.9 | 15.1 | N/A |
| | | 180 | 0 | 0 | 0 |
| ethyl methyl sulfide | C−C−S−C | 0 | 13.6 | 18.8 | N/A |
| | | 60 | 1 | 2.3 | N/A |
| | | 120 | 7.8 | 7.6 | N/A |
| | | 180 | 0 | 0 | N/A |

***Table 5.*** Optimized Nonbonded Parameters ($\epsilon_{ij}$ and $\delta_{ij}$) for Interactions between Small Particles

| interaction type | $\epsilon_{ii}$ (kJ/mol) | $\delta_{ii}$ (nm) |
|---|---|---|
| >**CH$_2$** | 0.45 | 0.390 |
| -**CH$_3$** | 1.00 | 0.390 |
| >**CH**- | 0.20 | 0.390 |
| =**CH**- | 0.45 | 0.375 |
| -**O**-, -**OH,** -CO-**OH** | 0.80 | 0.290 |
| >**C**=O, -**C**OO$^-$ | 0.25 | 0.360 |
| >C=**O,** -C**OO**$^-$ | 0.80 | 0.280 |
| N[a] | 1.00 | 0.330 |
| **-S-**, -**SH** | 2.50 | 0.340 |

[a] N indicates all types of nitrogen-containing groups.

***Table 6.*** Comparison of Densities and Self-Solvation Free Energies between Experiments and Calculations

| | density (g/cm$^3$) | | self-SFE (kJ/mol) | |
|---|---|---|---|---|
| pure liquid | cal.[a] | exp.[b] | cal. | exp.[c] |
| cyclohexane | 0.776 | 0.779 | −17.8 | −18.5 |
| n-pentane | 0.625 | 0.626 | −13.6 | −14.2 |
| isopentane | 0.635 | 0.620 | −13.5 | −13.5 |
| benzene | 0.848 | 0.879 | −17.9 | −19.1 |
| diethyl ether | 0.691 | 0.713 | −14.6 | −14.4 |
| 2,3-dimethyl-2-butene[d] | 0.659 | 0.708 | −20.2 | −17.7 |
| triethylamine | 0.764 | 0.726 | −18.9 | −18.8 |
| dimethylsulfide[e] | 0.816 | 0.846 | −16.0 | −15.7 |

[a] All the calculated values are obtained at 300 K. [b] Reference 64. [c] Measured at 298.15 K.[65] [d] The experimental self-solvation free energy of 2,3-dimethyl-2-butene was converted from the experimental vapor pressure of 2,3-dimethyl-2-butene at 298 K.[66] [e] The experimental self-solvation free energy of dimethylsulfide was converted from the experimental vapor pressure of dimethylsulfide at 293 K.[67]

sulfur in thiol and sulfide were obtained through simulations of pure dimethylsulfide.

As shown in Table 6, our parameters of small-small particle interactions reproduced the densities of eight pure

liquids with an average error about 3.2%, and the self-SFEs of these molecules had an average deviation of about 0.7 kJ/mol.

**3.4. Optimization of Nonbonded Parameters between Protein Particles and CG Water Particles.** One possible way to optimize the interaction parameters between protein particles and CG water particles could be to generate $\epsilon_{ij}$ and $\delta_{ij}$ from $\epsilon_{ii}$ and $\delta_{ii}$ of protein particles and CG water particles according to the LB combination rules. However, the LB rules are appropriate only if $\epsilon_{ii}$ and $\delta_{ii}$ are both associated with vdW interactions. The previously derived $\epsilon_{ii}$ for CG water reflects the overall effect of both vdW and electrostatic interactions between groups of four water molecules. Therefore, the LB rule cannot be fully used to obtain the $\epsilon_{ij}$ and $\delta_{ij}$ for interactions between protein particles and CG water particles. One the other hand, it would be useful to utilize the LB rules partially to simplify the optimization procedure. Thus, we adopted an optimization procedure in which (1) we utilized the LB rule to generate most of the $\delta_{ij}$, assuming that the size of the cavity created by a small particle in CG water is the vdW volume of the particle and (2) the $\epsilon_{ij}$ was obtained by reproducing the hydration free energies (HFE) of various classes of organic molecules. In our optimization procedure, the HFEs of 35 small molecules covering 11 classes of organic compounds, including alkanes, alcohols, ketones/aldehydes, ethers, carboxylic acids, amines, amides, aromatics, sulfides/thiols, alkyl ammoniums and carboxylate ions (Table 7), were used to determine the parameters of interactions between protein particles and CG water particles (Table 8).

**Alkanes.** All the $\delta_{ij}$ for interactions between alkyl particles and CG water particles were generated by the LB rules. The $\epsilon_{ij}$ for **$CH_4$** - W interactions was obtained by fitting the HFEs of methane. The $\epsilon_{ij}$ values for -**$CH_3$** -W, >**$CH_2$** -W, and >**CH**- -W were optimized by fitting the HFEs of ethane, propane, and isobutane simultaneously. As shown in Table 9, these parameters can accurately reproduce HFEs of the other ten alkane molecules that have linear, branched, and cyclic shapes. The average deviation of the calculated HFEs from experimental values was about 1.0 kJ/mol for these ten molecules and about 0.7 kJ/mol for all 14 alkane compounds.

**Alcohols and Ethers.** Methanol, ethanol, and 1-propanol were used to optimize the parameters for W and -**OH** interactions. During the optimization, we found that if the $\delta_{ij}$ of the W- -**OH** interaction was taken as 0.375 nm according to the combination rule, the magnitude of the HFEs of alcohols dropped too rapidly with an increasing length of the hydrocarbon chain. This trend does not agree with experimental data. We found that this problem can be alleviated by reducing the $\delta_{ij}$ for the W- -**OH** interaction to 0.280 nm. Our observations may be understandable since a polar -**OH** group has hydrogen bonds with individual water molecules and four water molecules in a CG water particle may not appear equivalent to an -**OH** group. Thus, the $\delta_{ii}$ for CG water, which is an average size of a cluster of four equivalent water molecules, may no longer be valid in generating the $\delta_{ij}$ for -**OH** - W interactions. Therefore, in our optimization for other polar particles, if a similar trend occurs, we optimize $\delta_{ij}$ instead of obtaining $\delta_{ij}$ according to the LB rules.

***Table 7.*** Experimental and Calculated Hydration Free Energies (kJ/mol) of Organic Compounds Used To Fit Parameters[a]

| compound | cal. | exp. |
|---|---|---|
| **Alkanes** | | |
| methane | 8.2 | 8.4[b] |
| ethane | 7.3 | 7.4 |
| propane | 8.2 | 8.3 |
| isobutane | 9.3 | 9.5 |
| **Alcohols** | | |
| methanol | −20.1 | −20.2[c] |
| ethanol | −20.5 | −21.0 |
| 1-propanol | −20.1 | −20.4 |
| **Ethers** | | |
| dimethyl ether | −7.7 | −8.0 |
| diethyl ether | −6.2 | −6.8 |
| **Amines** | | |
| methyl amine | −18.3 | −19.1 |
| ethyl amine | −19.4 | −18.8 |
| dimethyl amine | −19.3 | −18.0 |
| diethyl amine | −17.4 | −17.0 |
| trimethyl amine | −14.8 | −13.6 |
| triethyl amine | −11.4 | −12.6 |
| **Ketones and Aldehydes** | | |
| acetone | −14.1 | −16.1 |
| butanone | −15.2 | −15.2 |
| **Amides** | | |
| N-methylacetamide | −41.0 | −42.1 |
| acetamide | −39.7 | −40.6 |
| N,N-dimethylacetamide | −34.9 | −35.6 |
| N-propylguanidine | −44.8 | −44.8 |
| **Carboxylic Acids** | | |
| acetic acid | −27.4 | −28.0 |
| propionic acid | −27.8 | −27.1 |
| **Aromatics** | | |
| benzene | 0.0 | −3.6 |
| toluene | −3.7 | −3.7 |
| p-cresol | −29.0 | −25.6 |
| pyridine | −20.1 | −19.7 |
| **Alkyl Ammoniums[d]** | | |
| methylammonium | −303 | −298 |
| ethylammonium | −285 | −286 |
| **Carboxylate Ions[d]** | | |
| acetate | −338 | −334 |
| propionate | −325 | −328 |
| **Thiols and Sulfides** | | |
| methanethiol | −5.7 | −5.2 |
| ethanethiol | −4.8 | −5.4 |
| methyl methyl sulfide | −8.5 | −6.4 |
| methyl ethyl sulfide | −5.6 | −6.2 |

[a] All our calculations were carried out at 300 K. [b] Reference 65. [c] Reference 69. [d] Reference 70.

We also noticed that the inductive effect of heteroatoms could polarize alkyl chains, especially the carbon center that connects to heteroatoms, which could have more favorable interactions with water than with nonpolarized carbons. Thus, we introduced an extra type of carbon (-**$CH_x$**-P with P for polar group) that replaced the alkyl particles connecting to the -**OH** group. The $\epsilon_{ij}$ for W- -**$CH_x$** was optimized to be 1.15 kJ/mol. In addition to the hydroxyl group, -**$CH_x$** was also used to replace the alkyl particles that directly connect

***Table 8.*** Optimized Nonbonded Parameters ($\epsilon_{ij}$ and $\delta_{ij}$) for Interactions between Protein Particles and CG Water

| particle types interacting with CG water | $\epsilon_{ij}$ (kJ/mol) | $\delta_{ij}$ (nm) |
|---|---|---|
| W[a] | 5.00 | 0.47 |
| **$CH_4$** | 1.60 | 0.430 |
| **-$CH_3$** | 1.30 | 0.430 |
| **>$CH_2$** | 0.86 | 0.430 |
| **>CH-** | 0.45 | 0.430 |
| **-OH** | 14.00 | 0.280 |
| **-O-** | 14.20 | 0.280 |
| **-$NH_2$** | 13.50 | 0.280 |
| **>NH** | 20.50 | 0.280 |
| **>N-** | 24.00 | 0.280 |
| **>C**=O | 0.80 | 0.415 |
| >C=**O** | 5.25 | 0.375 |
| -CO-**NH**- | 5.10 | 0.400 |
| -CO-**$NH_2$** | 4.50 | 0.400 |
| -CO-**N**< | 4.50 | 0.400 |
| -CO-**OH** | 5.50 | 0.280 |
| **=CH-** | 0.78 | 0.4225 |
| **=N-** | 3.30 | 0.400 |
| **=NH** | 1.00 | 0.400 |
| **-C$OO^-$** | 2.00 | 0.415 |
| **-C$OO^-$** | 25.20 | 0.375 |
| **-$NH_3^+$** | 36.00 | 0.400 |
| **-SH** | 3.40 | 0.405 |
| **-S-** | 3.90 | 0.405 |

[a] The $\epsilon_{ii}$ and $\delta_{ii}$ of W−W interaction from the original model by Marrink et al.[18]

to ether, ketone, aldehyde, carboxylic acid, amine, amide, ammonium, and carboxylate ion groups.

The parameters for ether oxygen (-**O**-) were optimized in a similar way as those of -**OH**. The $\epsilon_{ij}$ for W- -**O**- was optimized by dimethyl ether and diethyl ether. Just like alcohols, $\delta_{ij}$ of the W- -**O**- is reduced to 0.28 nm.

**Amines.** There are three types of CG particles for amines. -**$NH_2$**, >**NH** and >**N**- are used to represent primary, secondary, and tertiary amines with its hydrogen, respectively. For each type, the simplest compounds (mono-, di-, or tri-) methyl and ethyl amine were used to optimize the parameters, respectively. Like $\delta_{ij}$ of W- -**OH** of alcohol, $\delta_{ij}$ of W- -**$NH_2$**, W- >**NH**, and W- >**N**- are all set to 0.28 nm to reduce the perturbation from neighboring carbon groups.

**Ketones, Aldehydes, Carboxylic Acids, and Amides.** Acetone and butanone were used to optimize the interaction parameters of carbonyl carbon (>**C**=O) and carbonyl oxygen (>C=**O**) with CG water. For simplicity, the aldehyde functional group (-**CH**=**O**) was considered to be composed of >**C**=O and >C=**O** particles. The parameters for carbonyl carbon and carbonyl oxygen were also applied to carboxylic acid and amide. Thus for amides, the parameters of interaction between amide nitrogen and CG water were next to be optimized. There are three kinds of amide nitrogens in our model: the N-monosubstituted (-CO-**NH**-) group that can be used for backbone amide, the -CO-**$NH_2$** group that can be used for side chains of Asn and Gln, and the N-disubstituted (-CO-**N**<) group. With the optimized carbonyl carbon and oxygen, we optimized parameters of the -CO-**OH** group of carboxylic acid by acetic acid and propionic acid.

**Aromatics.** There are two types of aromatic carbon groups. One carries a hydrogen atom and the other carries no hydrogen. We used =**CH**- to represent both of them. The HFEs of benzene, toluene, and p-cresol were used to optimize the parameters for the W- =**CH**- interaction. There are also two types of nitrogen groups on heterocyclic aromatic rings (=**N**- and >**NH**). The one (>**NH**) with hydrogen is a HB donor and the other one (=**N**-) is a HB acceptor. They cannot be represented by just one particle type. Therefore, the parameter for W- =**N**- interaction was obtained based on pyridine. The parameters for heterocyclic >**NH** are presumed to be the same as monosubstituted amide >**NH**. The average absolute deviation of the calculated solvation free energies of all aromatic compounds is 2.59 kJ/mol.

**Guanidine.** As there are no experimental solvation data for a series of guanidine compounds, we could not systematically optimize the parameters of different nitrogens for Arg side-chain analogue, N-propylguanidine. A guanidine group has three $sp^2$ nitrogens that connect to a central $sp^2$ carbon. Two of them have single bonds with the central carbon and the other one has a double bond. We assumed that the nitrogens that have single bonds have the same parameters ($\epsilon_{ij}$, $\delta_{ij}$) as amides (-**NH**-, -**$NH_2$**). In addition, the parameters of carbonyl carbon (>**C**=O) are used for central carbon in the guanidine group. Then, we optimized the interaction parameters for the nitrogen (=**NH**) that has a double bond.

All the optimized $\epsilon_{ij}$ and $\delta_{ij}$ are given in Table 8. The optimized $\epsilon_{ij}$ of interactions between all heteroatom groups and CG water were much larger than that for carbons. This could be understood since the $\epsilon_{ij}$ of these polar groups accounted not only for the vdW interactions with water but also the strong electrostatic interactions with water.

**3.5. Calculated Hydration Free Energies of Small Molecules.** These parameters allowed us to reproduce experimental HFEs of 105 compounds with an average absolute error of about 1.4 kJ/mol (Tables 7 and 9). In particular, we used 35 compounds as a fitting set to obtain all the parameters (Table 7). The average absolute error for the fitting set is 1.1 kJ/mol. For another 70 compounds as a test set (Table 9), the average absolute error is 1.6 kJ/mol. In addition, the correlation plot between the calculated and experimental HFEs is given in Figure 1. The square of the correlation coefficient ($R^2$) is 0.9993. Our parameters can reproduce HFEs with reasonable accuracy.

**3.6. Comparison of Transfer Free Energies of Side-Chain Analogues of Amino Acids from Cyclohexane to Water.** Several calculations use the transfer free energy of side-chain analogues of amino acids from cyclohexane to water to benchmark the accuracy of all-atom force fields.[48−50] With the optimized bonded and nonbonded parameters in hand, we were also able to compute the SFEs of the analogues in these two media and compare the results from experiments and the all-atom simulations (Table 10).

In our calculations, CG water was used to calculate SFEs in water and six >**$CH_2$** particles were used to build cyclohexane. Here, we calculated 19 amino acid side-chain analogues except for Gly and compared the results with the works by Shirts et al.,[49] Chang et al.,[50] MacCallum and Tieleman,[48] where 14−18 amino acid side-chain analogues were calculated. The analogue for Gly was dihydrogen,[39,40] which was out of the scope of our current CG model. The calculated SFEs of the analogues in water

***Table 9.*** Experimental and Calculated Hydration Free Energies (kJ/mol) of Organic Compounds Used To Test Parameters

| compound | cal. | exp. | compound | cal. | exp. |
|---|---|---|---|---|---|
| Alkanes | | | Aromatics | | |
| n-butane | 9.2 | 9.0 | o-xylene | −4.3 | −3.8 |
| n-pentane | 10.2 | 9.8 | naphthalene | −7.2 | −10.0 |
| isopentane | 9.8 | 10.0 | 2-methylpyridine | −22.0 | −19.4 |
| n-hexane | 10.9 | 10.7 | 3-methylpyridine | −22.9 | −20.0 |
| isohexane | 11.6 | 10.6 | methylindole | −31.0 | −24.3 |
| 3-methylpentane | 11.9 | 10.5 | methylimidazole | −39.9 | −42.1 |
| cyclohexane | 6.4 | 5.2 | Ethers | | |
| n-heptane | 11.4 | 11.0 | methyl isopropyl ether | −8.5 | −8.4 |
| methylcyclohexane | 10.9 | 7.2 | methyl n-propyl ether | −7.4 | −7.0 |
| n-octane | 13.3 | 12.1 | ethyl n-propyl ether | −6.0 | −7.6 |
| Alcohols | | | di-n-propyl ether | −3.2 | −4.8 |
| 2-propanol | −22.0 | −19.9 | methyl t-butyl ether | −7.8 | −9.3 |
| 1-butanol | −18.9 | −19.8 | di-isopropyl ether | −8.0 | −2.2 |
| 2-butanol | −19.4 | −19.1 | di-n-butyl ether | −1.9 | −3.5 |
| 2-methyl-1-propanol | −17.7 | −18.9 | tetrahydropyran | −10.7 | −13.1 |
| t-butanol | −23.9 | −18.9 | Ketones and Aldehydes | | |
| 1-pentanol | −16.7 | −18.7 | pentan-2-one | −13.9 | −14.8 |
| 2-pentanol | −17.7 | −18.4 | pentan-3-one | −14.9 | −14.3 |
| 3-pentanol | −17.0 | −18.2 | 3-methyl-2-butanone | −15.9 | −13.6 |
| 3-methyl-1-butanol | −17.5 | −18.5 | 2-hexanone | −12.2 | −13.8 |
| 2-methyl-2-butanol | −19.6 | −18.5 | 4-methyl-2-pentanone | −11.9 | −12.8 |
| 1-hexanol | −16.6 | −18.2 | 2-heptanone | −11.1 | −12.7 |
| 3-hexanol | −15.5 | −17.1 | 4-heptanone | −11.9 | −12.3 |
| 4-methyl-2-pentanol | −17.4 | −15.7 | 2,4-dimethyl-3-pentanone | −15.7 | −11.5 |
| 2-methyl-3-pentanol | −15.1 | −16.3 | acetaldehyde | −15.8 | −14.6 |
| 2-methyl-2-pentanol | −17.7 | −16.4 | propanal | −16.9 | −14.4 |
| cyclohexanol | −19.9 | −22.9 | butanal | −15.5 | −13.3 |
| 2,3-dimethyl-2-butanol | −16.3 | −16.4 | pentanal | −14.1 | −12.7 |
| Amines | | | hexanal | −12.8 | −11.8 |
| propyl amine | −18.0 | −18.4 | Thiols and Sulfides | | |
| butyl amine | −16.7 | −18.0 | propanethiol | −4.0 | −4.6 |
| pentyl amine | −16.0 | −17.2 | ethyl ethyl sulfide | −3.9 | −6.0 |
| hexyl amine | −15.4 | −16.9 | Carboxylic Acids | | |
| dipropyl amine | −14.8 | −15.3 | butyric acid | −26.6 | −26.6 |
| dibutyl amine | −11.0 | −13.9 | Alkyl Ammoniums | | |
| piperidine | −21.8 | −21.4 | propylammonium | −277 | −279 |
| N-methylpiperidine | −14.9 | −16.3 | butylammonium | −274 | −277 |
| Amides | | | Carboxylate Ion | | |
| propionamide | −39.2 | −39.3 | butyrate | −321 | −324 |
| N-acetylpyrrolidine | −37.1 | −40.9 | | | |

from the three all-atom simulation studies deviated on average from experimental values by about 3.4, 2.9 and 4.5 kJ/mol, respectively. Taking advantage of the simulation speed of our CG model, we were able to directly optimize our parameters by fitting the SFEs in water. As a result, the average deviation of our simulations was only about 1.3 kJ/mol.

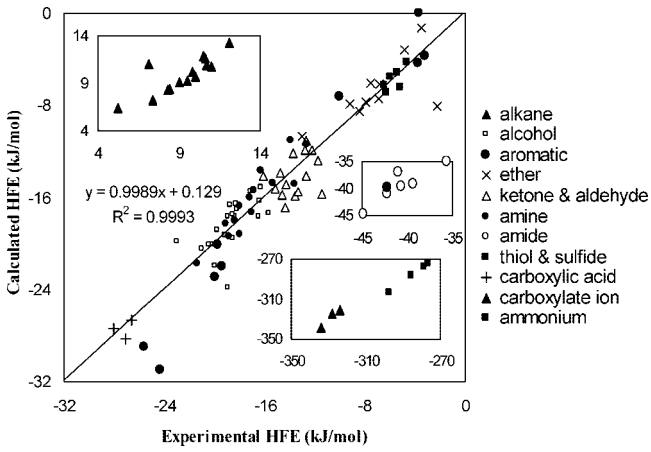



***Figure 1.*** Correlation plot between the calculated and experimental HFEs.

For the SFEs of these amino acid side-chain analogues in cyclohexane, our simulations had an average deviation of about 1.8 kJ/mol from experimental data. This is comparable to the all-atom simulations reported by Chang et al. (2.3 kJ/mol) and by MacCallum et al. (2.4 kJ/mol). Since the SFEs were all determined by interactions between protein particles, our result indicated that most of the assumptions made in the optimization of protein−protein particle interactions might work. However, there were a few exceptions. First, our model systematically underestimated the SFE of hydroxyl-containing molecules (Ser and Thr) by about 3 kJ/mol. Interestingly, even the all-atom model has the same problem, which was pointed out by Chang et al.[50] Second, the $\epsilon_{ij}$ of $>\mathbf{CH_2}$ - $=\mathbf{CH}$- interaction from the combination rule was so strong that the SFE of toluene could be exaggerated by more than 15 kJ/mol. The reason may be that our parameters for aromatic carbon were optimized from the self-SFE of pure benzene. In liquid benzene, the

***Table 10.*** Solvation Free Energies of Side-Chain Analogues of Amino Acids in Water or Cyclohexane from Our Calculation, the Simulations by Shirts et al.,[49] by Chang et al.,[50] and by MacCallum and Tieleman[48] Are Compared with Experimental Data[39,40]

| | SFE of side-chain analogues in water (kJ/mol) | | | | SFE of side-chain analogues in cyclohexane (kJ/mol) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | this work | Shirts | Chang | MacCallum | exp. | this work | Chang | MacCallum | exp. |
| Ala (methane) | 8.4 | 9.0 | 9.2 | 9.8 | 8.4 | 0.6 | 1.0 | 0.9 | 0.6 |
| Val (propane) | 8.2 | 10.7 | 10.4 | 12.0 | 8.3 | −8.7 | −6.6 | −6.5 | −8.6 |
| Leu (isobutane) | 9.3 | 11.3 | 11.1 | 13.7 | 9.5 | −14.0 | −9.2 | −10.3 | −11.0 |
| Ile (butane) | 9.2 | 11.2 | 10.5 | 12.2 | 9.0 | −11.8 | −9.9 | −8.8 | −11.6 |
| Ser (methanol) | −20.7 | −19.7 | −19.1 | −20.2 | −21.2 | −3.4 | −3.2 | −4.2 | −6.9 |
| Thr (propanol) | −20.1 | −18.7 | −19.4 | −20.3 | −20.4 | −6.7 | −6.6 | −7.2 | −9.7 |
| Cys (methanethiol) | −6.4 | −1.8 | −1.3 | −0.5 | −5.2 | −9.9 | −8.9 | −9.2 | −10.5 |
| Met (methylethylsulfide) | −6.9 | −0.3 | 0.1 | −7.1 | −6.2 | −16.7 | −14.5 | −14.2 | −16.0 |
| Asp (acetic acid) | −29.5 | | | −30.5 | −28.0 | −11.7 | | −13.1 | −9.2 |
| Glu (propionic acid) | −28.3 | | | −19.0 | −27.1 | −13.5 | | −16.6 | −13.9 |
| Lys (butylamine) | −16.7 | | | −13.6 | −17.9 | −15.0 | | −13.9 | −16.4 |
| Arg (N-propylguanidine) | −44.0 | | | −43.9 | −44.8 | −26.2 | | −22.7 | −20.6 |
| Asn (acetamide) | −40.4 | −35.6 | −35.7 | −34.5 | −40.5 | −12.9 | −13.7 | −12.9 | −14.3 |
| Gln (propionamide) | −39.2 | −35.5 | −36.6 | −31.4 | −39.2 | −15.0 | −16.8 | −17.7 | −16.1 |
| Pro (N-acetylpyrrolidine) | −37.1 | | | | −40.9 | −27.2 | | | −28.2 |
| Phe (toluene) | −3.7 | −2.7 | −2.9 | −1.2 | −3.2 | −23.0 | −19.7 | −20.2 | −17.5 |
| Tyr (p-cresol) | −29.0 | −21.2 | −20.5 | −18.8 | −25.5 | −24.9 | −22.7 | −23.0 | −25.0 |
| His (methylimidazole) | −39.9 | −36.2 | −35.6 | −28.0 | −42.9 | −21.7 | −19.0 | −19.7 | −23.4 |
| Trp (methylindole) | −31.0 | −17.3 | −24.0 | −16.2 | −24.6 | −34.1 | −38.1 | −30.6 | −34.3 |
| average error (kJ/mol) | 1.3 | 3.4 | 2.9 | 4.5 | | 1.8 | 2.3 | 2.4 | |

***Table 11.*** Transfer Free Energies of Side-Chain Analogues of Amino Acids from Cyclohexane to Water from Our Calculations, the Simulations by Chang et al.[50] and by MacCallum and Tieleman,[48] along with Those Determined by Experiments[39,40]

| | this work | Chang | MacCallum | exp. |
|---|---|---|---|---|
| Ala (methane) | 7.8 | 8.2 | 8.9 | 7.8 |
| Val (propane) | 16.9 | 17.0 | 18.5 | 16.9 |
| Leu (isobutane) | 23.3 | 20.3 | 24.0 | 20.5 |
| Ile (butane) | 21.0 | 20.4 | 21.0 | 20.6 |
| Ser (methanol) | −17.3 | −15.9 | −16.0 | −14.3 |
| Thr (propanol) | −13.4 | −12.8 | −13.1 | −10.7 |
| Cys (methanethiol) | 3.5 | 7.6 | 8.7 | 5.3 |
| Met (methylethylsulfide) | 9.8 | 14.6 | 7.1 | 9.8 |
| Asp (acetic acid) | −17.8 | | −17.4 | −18.8 |
| Glu (propionic acid) | −14.8 | | −2.4 | −13.2 |
| Lys (butylamine) | −1.7 | | 0.3 | −1.5 |
| Arg (N-propylguanidine) | −17.8 | | −21.2 | −24.2 |
| Asn (acetamide) | −27.5 | −22.0 | −21.6 | −26.2 |
| Gln (propionamide) | −24.2 | −19.8 | −13.7 | −23.1 |
| Pro (N-acetylpyrrolidine) | −9.9 | | | −12.7 |
| Phe (toluene) | 19.3 | 16.8 | 19.0 | 14.3 |
| Tyr (p-cresol) | −4.1 | 2.2 | 4.2 | −0.5 |
| His (methylimidazole) | −18.2 | −16.6 | −8.3 | −19.5 |
| Trp (methylindole) | 3.1 | 14.1 | 14.4 | 9.7 |
| average error (kJ/mol) | 2.2 | 2.4 | 4.2 | |

aromatic−aromatic interactions and/or the electrostatic interactions are always present between benzene molecules, and they therefore contribute to the optimized values for the **=CH-** - **=CH-** parameters. But no such interactions exist between **>CH$_2$** and **=CH-** particles. The same exaggeration, though not very large, could also be seen in the all-atom models.[50] Thus, the $\epsilon_{ij}$ for **>CH$_2$** - **=CH-** interaction was separately optimized. It is 0.34 kJ/mol.

Finally, the transfer free energies of the side-chain analogues from cyclohexane to water were calculated (Table 11). The transfer free energies from our model agree reasonably well with experimental data with an average deviation of 2.2 kJ/mol. These results are comparable to those obtained by all-atom simulations reported by Chang et al. and MacCallum et al., which have average absolute errors of 2.4 and 4.2 kJ/mol, respectively.

**3.7. Limitation.** We treat the electrostatic interaction in a similar way to that of the CG model of Marrink et al. by implicitly incorporating electrostatic terms into vdW terms. Thus, the long-range electrostatic interaction beyond the cutoff (1.2 nm) is missing. As pointed out before,[18,20] the results for a system where the long-range electrostatics are important may be affected. For a system such as a huge protein that has electrostatic complex structures in its low-dielectrics interior, the results with our model should be interpreted with caution.

## 4. Conclusion

Further development and parametrization of our coarse-grained protein model in tandem with a CG solvent model have been carried out. Totally 24 types of protein particles were built to describe various amino acids. The interaction parameters between protein particles and between protein and CG water particles were optimized based on the self-solvation free energies of eight molecules and the hydration free energies of a set of 35 molecules. Our model can reproduce hydration free energies of 105 organic compounds with an average deviation of about 1.4 kJ/mol from experimental data. It can also give transfer free energies of the side-chain analogues of all 19 amino acids from cyclohexane to water with an average absolute error of about 2.2 kJ/mol. The results showed that parametrization based on solvation properties enables our model to capture solvation changes in amino acids, which is essential for protein folding. It is possible to incorporate our protein model systematically into a CG solvent model for effective studies of protein folding and protein−protein interactions. Full parametrization in this direction is in progress.

**Acknowledgment.** We are grateful to RGCHK (N-HKUST 623/04, CA06/07.SC05) and NSFC (20225312) for financial support of the research.

**Supporting Information Available:** Schematic views of optimized geometries of small molecules by quantum mechanics calculation (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

## References


(1) Urbanc, B.; Borreguero, J. M.; Cruz, L.; Stanley, H. E. *Methods Enzymol.* **2006**, *412*, 314–338.

(2) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147.

(3) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860.

(4) Smith, A. V.; Hall, C. K. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 344.

(5) Marchut, A. J.; Hall, C. K. *Proteins* **2007**, *66*, 96–109.

(6) Akkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure* **2006**, *14*, 1767.

(7) Gô, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183.

(8) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844–11849.

(9) Thirumalai, D.; Guo, Z. *Biopolymers* **1995**, *35*, 137.

(10) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616–11629.

(11) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394–408.

(12) Sharma, S.; Ding, F.; Dokholyan, N. V. *Biophys. J.* **2007**, *92*, 1457–1470.

(13) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.

(14) Izvekov, S.; Voth, G. A. *J. Chem. Theor. Comput.* **2006**, *2*, 637.

(15) Ayton, G. S.; Voth, G. A. *J. Struct. Biol.* **2007**, *157*, 570.

(16) Shi, Q.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 15045.

(17) Feig, M.; Brooks, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(18) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.

(19) Monticelli, L.; Kandasamy, S.; Periole, X.; Larson, R.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(20) Marrink, S. J.; Risselada, H, J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys Chem. B* **2007**, *111*, 7812.

(21) Baron, R.; Trzesniak, D.; de Vries, A. H.; Elsener, A.; Marrink, S. J.; van Gunsteren, W. F. *Chemphyschem* **2007**, *8*, 452.

(22) Kasson, M. P.; Kelly, N. W.; Singhal, N.; Vrljic, M.; Brunger, A. T.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11916.

(23) Marrink, S. J.; Mark, A. E. *J. Am. Chem. Soc.* **2003**, *125*, 15233.

(24) de Vries, A. H.; Mark, A. E.; Marirnk, S. J. *J. Am. Chem. Soc.* **2004**, *126*, 4488–4489.

(25) de Vries, A. H.; Yefimov, S.; Mark, A. E.; Marrink, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5392.

(26) Bond, P. J.; Sansom, M. S. P. *J. Am. Chem. Soc.* **2006**, *128*, 2697.

(27) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.

(28) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. *J. Am. Chem. Soc.* **2007**, *129*, 10126.

(29) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Sligar, S. G.; Schulten, K. *J. Phys. Chem. B* **2007**, *111*, 11095.

(30) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593.

(31) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. P. *Structure* **2008**, *16*, 621.

(32) Sansom, M. S. P.; Scott, K. A.; Bond, P. J. *Biochem. Soc. Trans.* **2008**, *36*, 27.

(33) Honig, B.; Yang, A. S. *Adv. Protein Chem.* **1995**, *46*, 27–58.

(34) Feig, M.; Brooks, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(35) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.

(36) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–7592.

(37) Fauchere, J.-L.; Pliska, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.

(38) Kim, A.; Szoka, F. C. *Pharm. Res.* **1992**, *9*, 504.

(39) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664.

(40) Wolfenden, R.; Anderson, L.; Cullis, P.; Southgate, C. *Biochemistry* **1981**, *20*, 849.

(41) Wimley, W. C.; Creamer, T. P.; White, S. H. *Biochemistry* **1996**, *35*, 5109.

(42) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638.

(43) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276.

(44) Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 569–1489.

(45) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(46) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(47) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656.

(48) MacCallum, J. L.; Tieleman, D. T. *J. Comput. Chem.* **2003**, *24*, 1930.

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

(50) Chang, J.; Lenhoff, A. M.; Sandler, S. I. *J. Phys. Chem. B* **2007**, *111*, 2098.

(51) Han, W.; Wu, Y.-D. *J. Comput. Theory Chem.* **2007**, *3*, 2146.

(52) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730–748.

(53) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.

(54) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.

(55) Berendsen, H. J. C.; Pstma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(56) Mezei, M.; Beveridge, D. L. *Ann. N. Y. Acad. Sci.* **1986**, *482*, 1.

(57) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Greber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529.

(58) Frisch, M. J. et al. *Gaussian 03, revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

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

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

(61) Kitagawa, T.; Miyazawa, T. *Bull. Chem. Soc. Jpn.* **1968**, *41*, 1976.

(62) Herrebout, W. A.; van der Veken, B. J.; Wang, A.; Durig, J. R. *J. Phys. Chem.* **1995**, *99*, 578.

(63) Henderson, D.; Leonard, P. J. Liquid Mixture. In *Physical Chemistry: An Advanced Treatise*; Academic Press: London, 1971; Vol *VIIIB*.

(64) Lide, D. R. In *CRC Handbook of Chemistry and Physics*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(65) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016.

(66) Schurman, I.; Boord, C. E. *J. Am. Chem. Soc.* **1933**, *55*, 4930.

(67) Osborne, D. W.; Doescher, R. N.; Yost, D. M. *J. Am. Chem. Soc.* **1942**, *64*, 169.

(68) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253.

(69) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563.

(70) Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4118.

CT800184C

JCTC Journal of Chemical Theory and Computation

# Ab Initio and Density Functional Theory Modeling of the Chiroptical Response of Glycine and Alanine in Solution Using Explicit Solvation and Molecular Dynamics

Matthew D. Kundrat and Jochen Autschbach*

*Department of Chemistry, 312 Natural Sciences Complex, The State University of New York at Buffalo, Buffalo, New York 14260-3000*

Received July 15, 2008

**Abstract:** We investigate ways in which simple point charge (SPC) water models can be used in place of more expensive quantum mechanical water molecules to efficiently model the solvent effect on a solute molecule's chiroptical responses. The effect that SPC waters have on the computed circular dichroism of a solvated glycine molecule are comparable to, albeit somewhat weaker than, that of quantum mechanical waters at the coupled cluster CC2 level of theory. The effects of SPC waters in fact correlate better with QM-CC2 waters than quantum mechanical waters computed with density functional theory (DFT) methods, since they do not promote spurious charge transfer excitations that are a known deficiency with most popular density functionals. Furthermore, the near zero order scaling of point charge waters allows multiple layers of explicit solvation to be modeled with negligible computational cost, which is not practical with CC2 or DFT levels. As a practical example, we model the molar rotations of glycine and alanine, and track their convergence.

## Introduction

Molecular modeling of molecules in solution poses a continuing challenge. The amino acids which comprise our proteins are of great interest, and modeling them in solution phase is drawing increasing interest. Glycine, being the smallest of this class of molecules, is a natural target for ab initio computations. Many recent papers have been published on the solvation of this molecule, several of which we cite here.[1−12] Alanine, the second smallest amino acid and the smallest which is chiral, has also gathered significant attention.[8,13−15]

A recent focus of our research has been modeling the chiroptical response properties of amino acids in solution.[16−19] These interactions of chiral molecules with polarized light include specific rotation/molar rotation at fixed wavelengths, optical rotatory dispersion, and circular dichroism.[20,21] We as well as others[8,22−24] have employed continuum solvation models[25] in our earlier studies of these properties. Such methods, among the most popular being the Polarizable Continuum Models (PCMs) and the COnductor like Screening MOdel (COSMO), do well to model bulk solvent effects. However they fare more poorly for short, explicit solvent-solute interactions such as hydrogen bonding, which is important for aqueous solutions of highly polar molecules such as glycine. For these, a more detailed solvent model is called for, such as one incorporating several solvent molecules explicitly. The exploration of such explicit solvation on chiroptical properties has only recently begun.[26−28] Here we continue this exploration with a look at how differing explicit solvation models affect the chiroptical response properties of the glycine and alanine molecules and at the performance of simplified models that, unlike the continuum models, treat the solvent molecules as discrete entities. Replacing a continuum such as COMSO or PCM by discrete solvent molecules comes at the added cost of averaging over many solvent-solute configurations. For this purpose, we employ molecular dynamics calculations.

This work begins with a look at how water molecules congregate around a glycine molecule that they are solvating and how those solvent effects decrease as a function of solute-solvent distance with various model systems. It next probes more deeply into how the solvent molecules are

* Corresponding author e-mail: jochena@buffalo.edu.

10.1021/ct8002767 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

perturbing the molar rotation and circular dichroism response of the glycine molecule, paying particular attention to how the solvent affects the lowest energy electronic excitation of the solute. It then explores the differences and similarities of using quantum mechanical water molecules and those comprised simply of point charges for solvation modeling, discussing the benefits and shortcomings of both methods. Finally the average molar rotations resulting from thousands of TDDFT calculations of glycine and alanine will be compared with experiment.

## Computational Methods

Many of the computational methods used in this work are detailed in two previous publications, where TDDFT based computations of optical rotations of amino acids were exhaustively benchmarked.[17,19] All quantum mechanical data were computed with the Turbomole[29] quantum chemical software, version 5.7.1. The calculations were performed with the Becke three parameter B3LYP and BHLYP hybrid functionals[30] as well as the CC2 coupled cluster method[31] as implemented in the Turbomole code. Dunning's aug-cc-pVDZ basis set[32] was used. For some calculations the COnductor-like Screening MOdel (COSMO)[33] of solvation was applied to the ground state. Molar rotations were calculated at the wavelength of the sodium D line (589.3 nm). All molar rotations are reported in units of deg·cm$^2$/(dmol). The center of mass of the glycine molecule has been used for the coordinate origin for all response calculations. While results attributed to the length representation of the electronic dipole operator are formally origin dependent, this dependence is minimized in variational methods when large basis sets such as aug-cc-pVDZ are used; see our earlier work and the references cited therein.[17] Origin dependence does not vanish upon basis set saturation for the CC2 method.[34,35] However, the small difference in the results using the origin dependent length operator and the independent velocity operator did not effect our conclusions.

Geometries used in the quantum mechanical calculations were generated with the GROMACS[36] molecular dynamics program, version 3.3.3, in a fashion similar to Mukhopadhyay et al.[26] Molecular dynamics of a glycine molecule were run in a cubic periodic solvent box measuring 25 × 25 × 25 nm containing 509 Simple Point Charge (SPC) water molecules, resulting in a system with an average density of 1.0 g/cc. To avoid possible artifacts from the cubic shape of the periodic box, no water molecules greater than 12.5 nm from the center of the solute were included in the subsequent quantum mechanical calculations. The all-atom OPLS-AA[37] force field was used for the simulations, which were carried out at 300 K with a time step of 1 fs. Snapshots of the simulations for subsequent computations of Circular Dichroism (CD) and optical rotation were taken every 10 ps, which was a sufficient duration for energetic and chiroptical response calculations of adjacent configurations to be uncorrelated.

## Results and Discussion

**Glycine Solvation As a Function of Distance.** Explicit solvation at the quantum mechanical level entails significant computational costs. With this in mind, any development of such a solvation model should first consider how many explicit solvent molecules are needed in a system in order to affect full solvation on the solute. Fortunately, with respect to the solvation of the amino acid glycine, we can benefit from the experience gained in previous studies. Several articles have been published on the topic, much of which is summarized in the recent work of Aikens and Gordon.[2] They concluded, inter alia, that "Eight water molecules do not appear to fully solvate the glycine molecule.". As we are interested in the effects that water molecules have on the chiroptical response properties of amino acids such as glycine, we must further investigate just how many of these molecules are needed to "fully solvate" glycine, so that we may have some idea of how many water molecules should be included in system that aims to model the full effects of explicit water molecules on this solute.

Before delving into the energetics of solvation, let us first consider the structure that the water molecules form around the glycine solute. To look at how solvent molecules orient themselves about the solute a radial distribution plot was created from 4096 snapshots of a glycine in water dynamics simulation. The results are depicted in Figure 1.

From the plot it is apparent that the density of the solvent reaches a maximum around 3.7 Å from the solute center and then dwindles to a minimum at 5.0 Å. Integration of the area between the solute and the minimum at 5.0 Å reveals approximately 16 water molecules are contained in this innermost solvation shell. Integration through the next minimum in solvent density at 7.0 Å adds another 30 water molecules in a second solvation sphere, giving a total of 46 waters needed to fill two solvation spheres. Beyond this the water appears to have little structure with respect to the solute and maintains a density close to that of pure water. If we set a cutoff for a third solvation shell at 9.0 Å, it would mean including approximately 100 water molecules in the first three shells. A typical snapshot of a molecular dynamics simulation of glycine including one, two, and three solvent shells of water is illustrated in Figure 2.

To have a uniform reference for the solvation energy, all energies for the interaction between the solvent and the solute were obtained from quantum mechanical calculations. In the molecular dynamics framework used here, such calculations are inherently more time-consuming than the dynamics calculations used to generate the configurations, so the number of water-glycine interactions as well as the number of sampling snapshots had to be limited. Here we considered at the individual interactions of each of the nearest 256 water molecules with the glycine solute. Since this entailed performing that number of costly ab initio calculations for every snapshot examined, only the first eight sampling geometries of the dynamics simulation were used for this example. In the interest of possibly saving computational resources in the future, each water molecule was modeled once in the full quantum mechanical (QM) set, with a full compliment of electrons and basis functions, and once as a set of negative (oxygen) and positive (hydrogen) point charges. The charges for this model were those from the 3 point

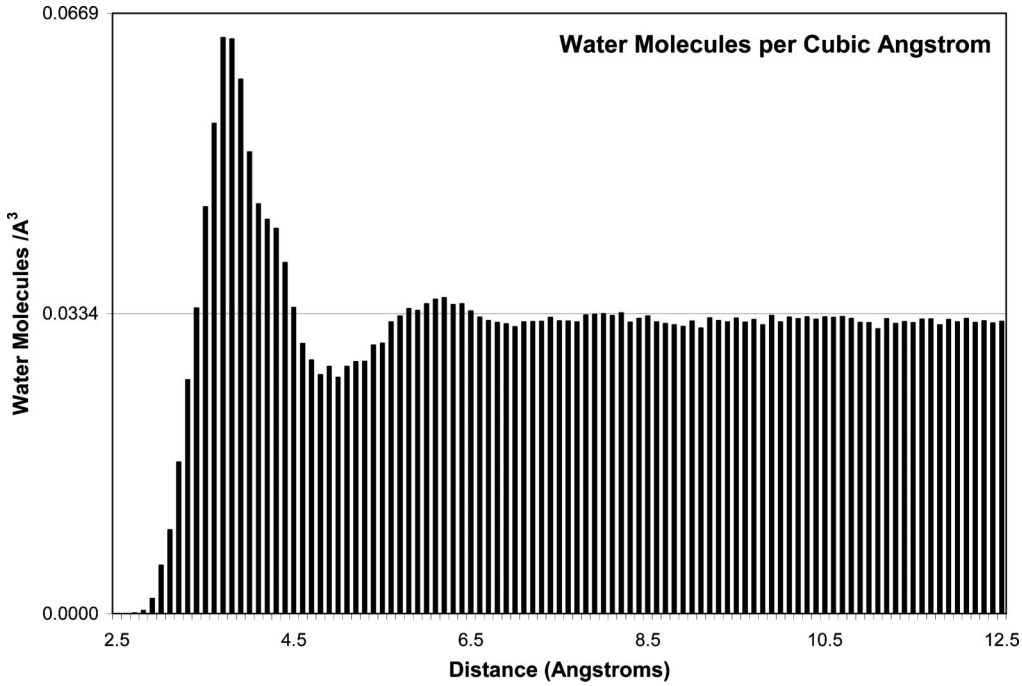



***Figure 1.*** Molecular dynamics simulation of glycine in water: Probability of finding a water molecule (per cubic angstrom) as a function of distance from the solute. Solute position is defined as the center of solute mass; solvent position is defined as the location of its oxygen nucleus. Radial shells with a width of 0.1 Å were used to generate the plot. For reference, pure water with a density of 1.0 g/cm$^3$ has 0.0334 water molecules per cubic angstrom.

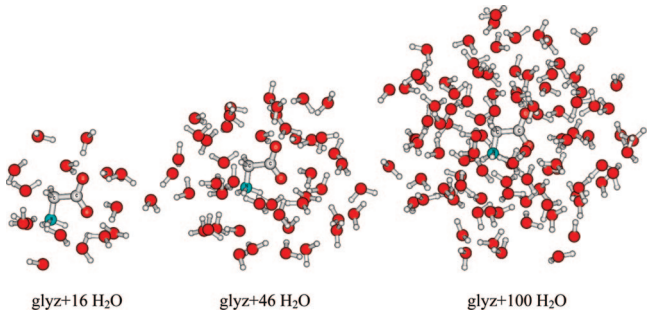



***Figure 2.*** Representative configurations of a glycine molecule surrounded by 16, 46, and 100 water molecules which approximate one, two, and three shells of solvation, respectively.

TIP3P type waters.[38] For each calculation the interaction energy was calculated by subtracting the sum of the energy of the glycine molecule and the water molecule from that of the glycine-water system (no correction was made for basis set superposition). Note that this is only really partial solvation energy, since only a single molecule of solvent is being considered. The results are depicted in Figure 3.

The energetic results are consistent with what is expected from the radial distribution of the water molecules. The water molecules in the innermost solvation shell, those within five angstroms of the solute center, are those with the greatest effect on the energy. This effect is largely negative, which indicates the stabilization of the highly polar glycine zwitterion by the polar water molecule. Note that this stabilizing effect is seen when point charge water molecules are used as well as when full QM waters are modeled. This indicates that the dipole-dipole interaction, which can be effectively modeled by the SPC waters, plays a prominent role in the solvation.

For water molecules farther from the solute, the solvation effect is lessened and becomes dominated by pure dipole−dipole interactions. The energetic effect of waters in the second solvation shell, those between approximately 5 and 7 Å, is far smaller, on average less than half the magnitude of those in the inner shell. Beyond that distance solvation effects dwindle further. At 9 Å from the solute center, about where the 100th water is located, the effect upon solvation has become small compared to the level of accuracy inherent in the ab initio calculations.

**Glycine: Solvent Effects on Molar Rotation.** The recent focus of our research has been the modeling of the chiroptical properties, including molar rotation, of amino acids in solution. Incorrect modeling of the Boltzmann populations of amino acid conformers, caused in part by an insufficient solvent system, has been considered as one possible source of error. This is especially true since these conformers can have both large positive and negative rotations, which should match experimental results only upon calculation of a correct weighted average of their responses. Another important part of this current model, which was absent in earlier works, is the direct effect that explicit solvent molecules have on optical rotation. It is important to study their effects separately, which makes the glycine and alanine zwitterions attractive candidates for study since each has only one local minimum structure.[17] Glycine draws particular interest not only because of its low molecular weight but also since it is symmetrical, making it a good test case as a system which can exhibit a chiroptical response depending on its

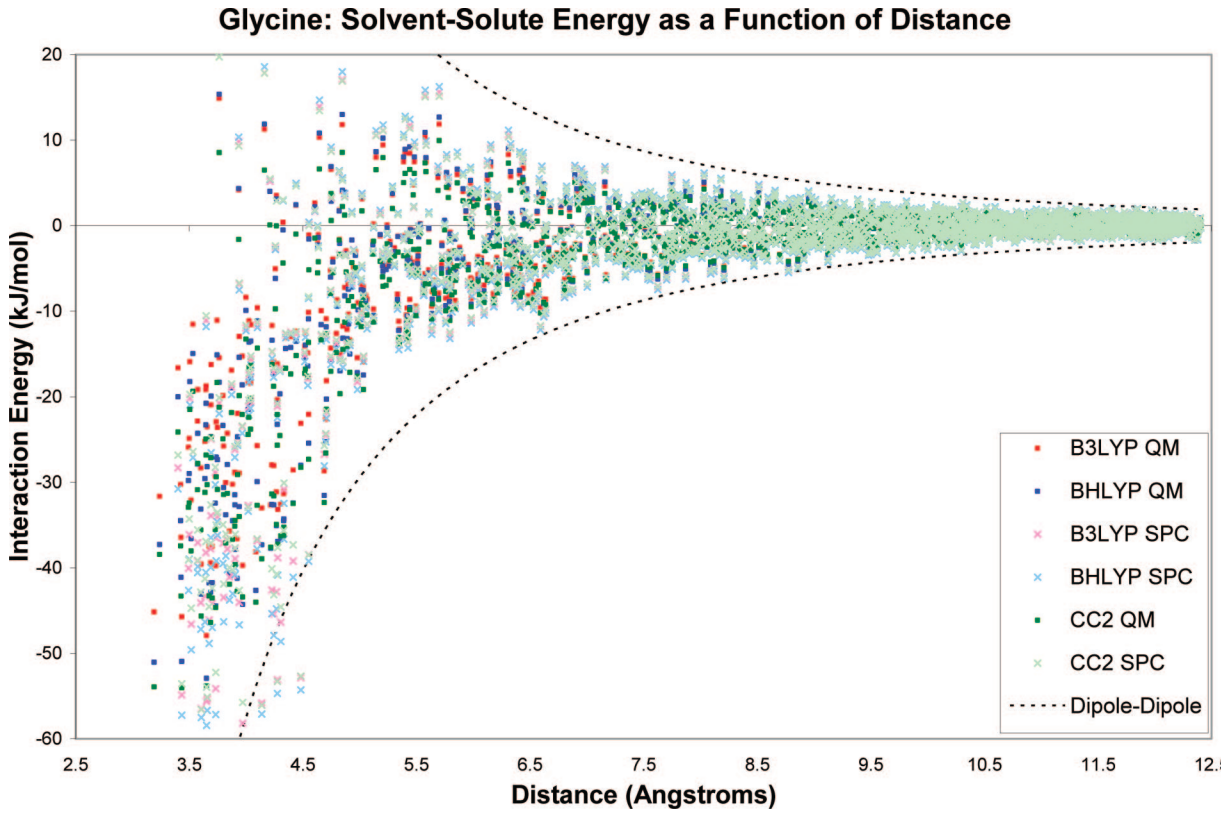



***Figure 3.*** The energy of solvation caused by one water molecule upon glycine as a function of the distance between the oxygen atom of the water and the center of mass of the glycine. ($E_{interaction} = E_{solute+solvent} - E_{solute}$, with $E_{solute} = 0$ when SPC waters are used.) The closest 256 water molecules considered at in each of eight snapshots of a molecular dynamics simulation. Calculations at the B3LYP, BHLYP, and CC2 are shown in red, blue, and green, respectively. Full quantum water results are shown in dark colors, while those from SPC point charge waters are in lighter colors. The range of variation in energy expected from static dipole-dipole interactions is depicted by the dashed line, computed using dipole moments of 13 and 2.35 Debye for glycine (using an average value from Destro et al. and references cited therein)[39] and water (using the value for the TIP3P model),[38]respectively.

configuration, responses which we know must average to zero over time.

To investigate the aspects of explicit solvation a linear response calculation of molar rotation at 589.3 nm of a glycine-water system was performed on each of the MD snapshots considered in the previous section. Again both full QM and SPC waters were used, and the B3LYP and BHLYP hybrid functionals were considered. The CC2 method was not used here because of the lack of a CC2 implementation for optical rotation in the software used for this work.

While the glycine molecule is itself an achiral molecule, a snapshot of a solvated glycine can exhibit a chiroptical response for a combination of two reasons: First, the glycine molecule itself can be found in geometries where its plane of symmetry is broken. It has an equal probability of being captured in both dextrorotatory and levorotatory conformations. For the limited sample size of eight configurations, we were fortunate to find that the unsolvated glycine molecule had a positive molar rotation exactly four times and an equal number of negative molar rotations. The second reason that a solvated glycine molecule may exhibit a nonzero molar rotation is the asymmetrical orientation of a water molecule with respect to the glycine. This is the phenomenon of interest; in order to isolate this solvent effect on molar rotation response calculations were carried out on glycine both with and without the presence of a water molecule and subtracting the two molar rotations. (If the water molecule could adopt a chiral configuration,[40] we would have to consider its molar rotation as well; however, an isolated water molecule has only three atoms which are inherently coplanar and thus can have no intrinsic chirality regardless of how its geometry is distorted by vibration.) The results of these calculations are shown in Figure 4.

As expected, a water molecule has a pronounced effect on the molar rotation of a glycine-water system, and the magnitude of that effect varies inversely with the distance between the solvent and solute molecules. The magnitude of this change in molar rotation remains rather consistent regardless of whether a point charge or quantum mechanical water causes the perturbation. A drastic change however can be seen when the amount of exact exchange in the hybrid DFT functionals is varied. The change in molar rotation caused by solvation is much greater in the B3LYP hybrid than in the BHLYP; the cause of this change will be discussed in the next section. The large magnitude of the solvent effect on the molar rotation illustrates the challenge of successfully averaging this quantity over the course of a

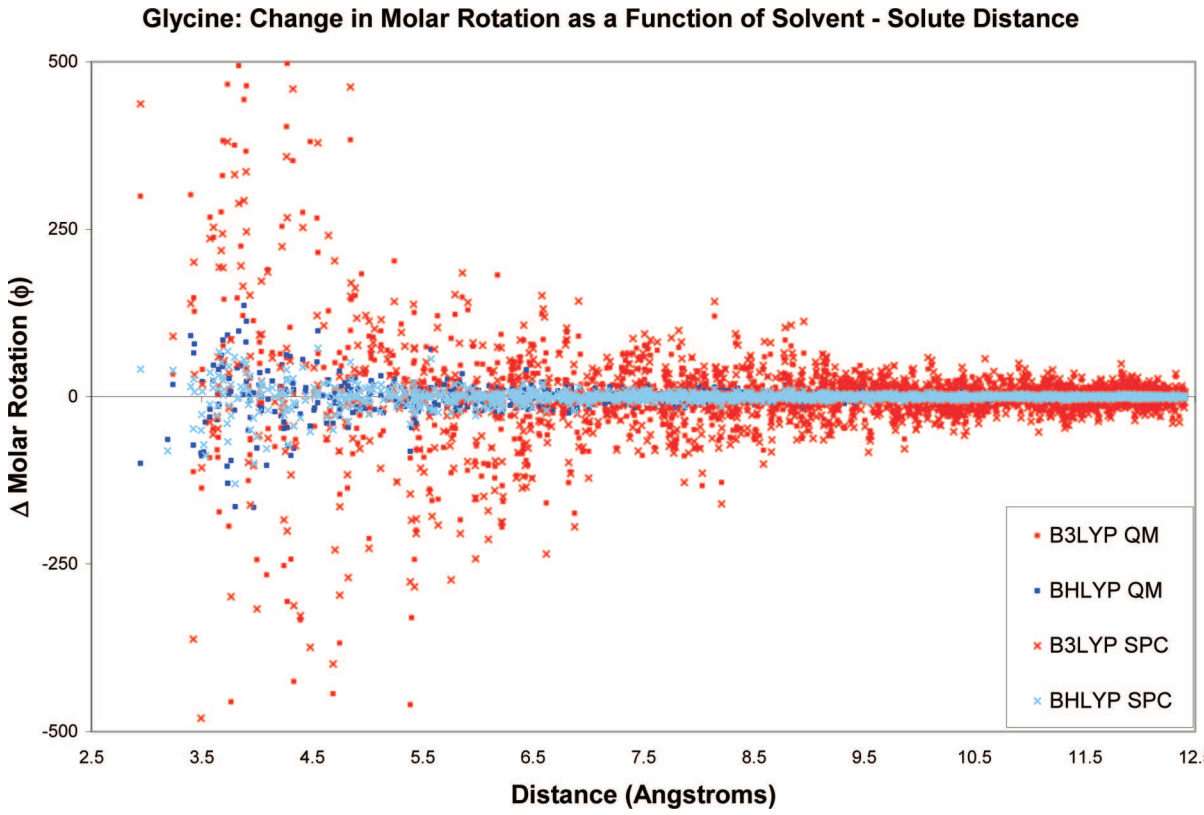



***Figure 4.*** The change in molar rotation ($\phi$) caused by a single water molecule as a function of its distance from the glycine center ($\Delta\phi = \phi_{\text{solute+solvent}} - \phi_{\text{solute}}$). Calculations were performed with the aug-cc-pVDZ basis at the B3LYP (red) and BHLYP (blue) levels of theory using QM (dark colored) and SPC (light colored) water molecules.

molecular dynamics simulation in order to minimize the statistical error.

**Effects on the Circular Dichroism of the Lowest Excitation.** The direct linear response method for calculating molar rotation used in the previous section performs the task very efficiently. However, it tells us little about *why* a molar rotation is what it is. Here we are interested in what electronic transitions are responsible for the molar rotation. This can be done by computing molar rotation via the sum over states method. In a prior study on aliphatic amino acids such as glycine, we confirmed that the lowest lying electronic excitation, the n to $\pi^*$ transition of the carboxylate moiety has a large influence on the molar rotation observed at 589.3 nm.[16] While to accurately model molar rotation by the SOS method nearly all possible excitations do need to be included, the excitation closest in energy to 589.3 nm is particularly important regarding the contribution from the carboxylate chromophor. As we are interested now in the effect an explicit solvent molecule has on this molar rotation, it is therefore beneficial to consider the effect this solvation has on the critical first electronic CD.

Using the same configurations as in the previous section, calculations of the rotatory strength and wavelength of the first excitation were performed at the B3LYP, BHLYP, and CC2 levels of theory using QM and SPC water molecules. In each case the change of the CD response caused by the solvent molecule was computed by subtracting from the CD of each water-glycine system the CD caused by an unsolvated glycine molecule in the same geometry. The lowest electronic excitation of water is far higher in energy than that of the glycine, and so the first CD transition that is observed in the model is always centered on the solute. The solvent serves merely to perturb that transition within the solute. The results of these calculations are plotted in Figure 5.

Just as with the energy and molar rotation, the effect of a water molecule on the first CD excitation decreases with distance, as expected. The change in rotatory strength caused by the water has an equal probability of being positive or negative, as one would expect for a chiral molecule being perturbed by an achiral solvent. The change in the wavelength of that transition however tends to be negative at all levels of theory. That is, a nearby water molecule tends to induce a blue shift of the first transition.

This blue shift is far more pronounced at the B3LYP and CC2 levels of theory than with the BHLYP functional. In the gas phase model, this excitation takes place at 430, 393, and 287 nm, respectively. These are all far too low in energy compared to solution phase measurements of aliphatic amino acid CD, which indicate that this dichroism has a maximum in the range of 200 to 215 nm.[41,42] Solvation increases this excitation energy and thus lowers its wavelength at all levels of theory, more so for B3LYP and CC2 than BHLYP. The contribution of a CD excitation to the molar rotation depends inversely on the difference between the wavelength at which the molar rotation is computed or measured (in this case that of the sodium D line, 589.3 nm) and the wavelength at which

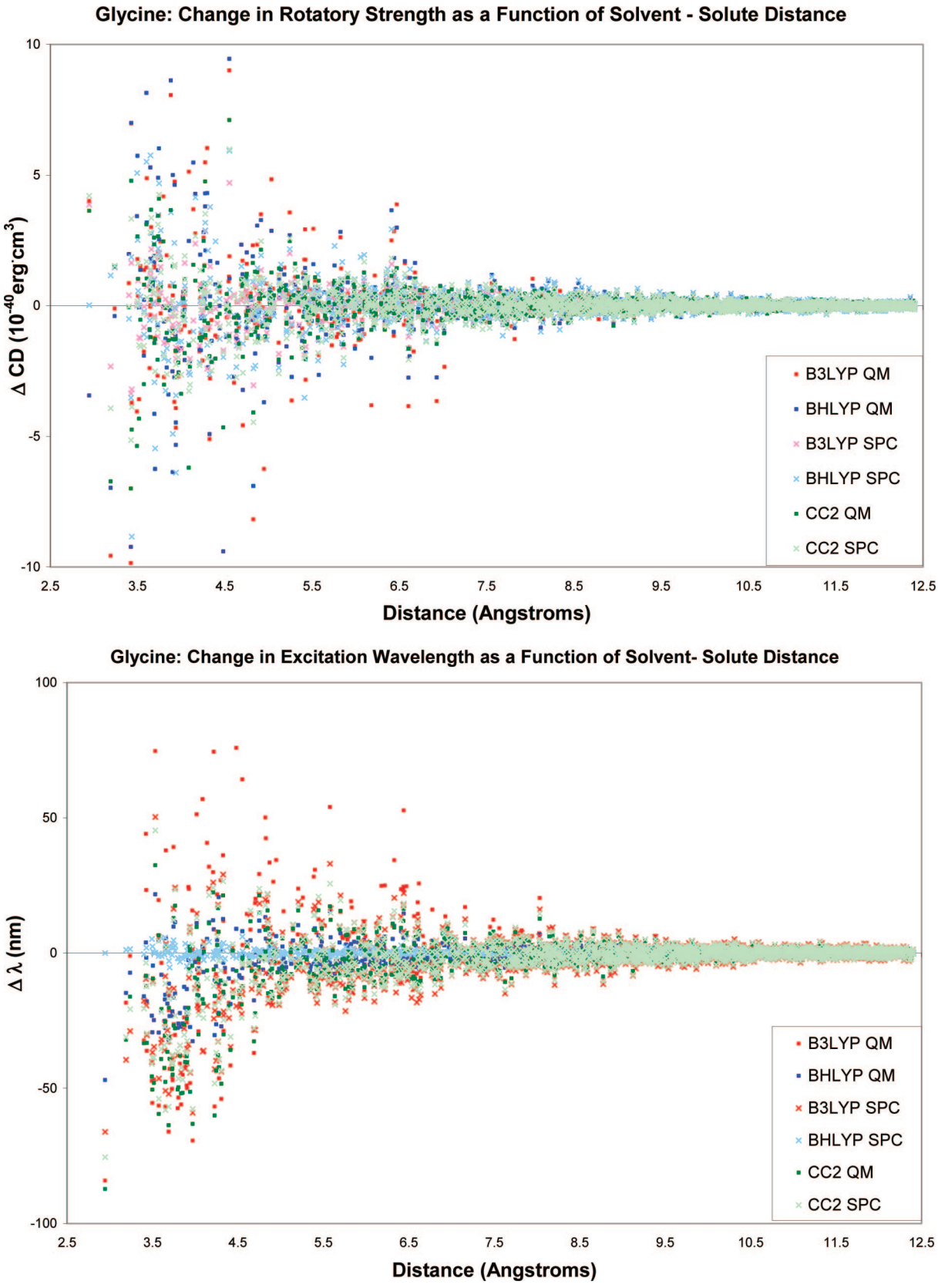



***Figure 5.*** The change in circular dichroism of the first electronic transition as a function of water distance from the solute center. Change in rotatory strength (top) and wavelength (bottom) are plotted.

the excitation occurs, per the Kramers−Kronig relationship. This leads to the explanation of why solvation has a more pronounced effect on the molar rotation of the glycine molecule with the B3LYP hybrid functional than with the BHLYP. Solvation shifts the first electronic excitation wavelength farther from the sodium D line to

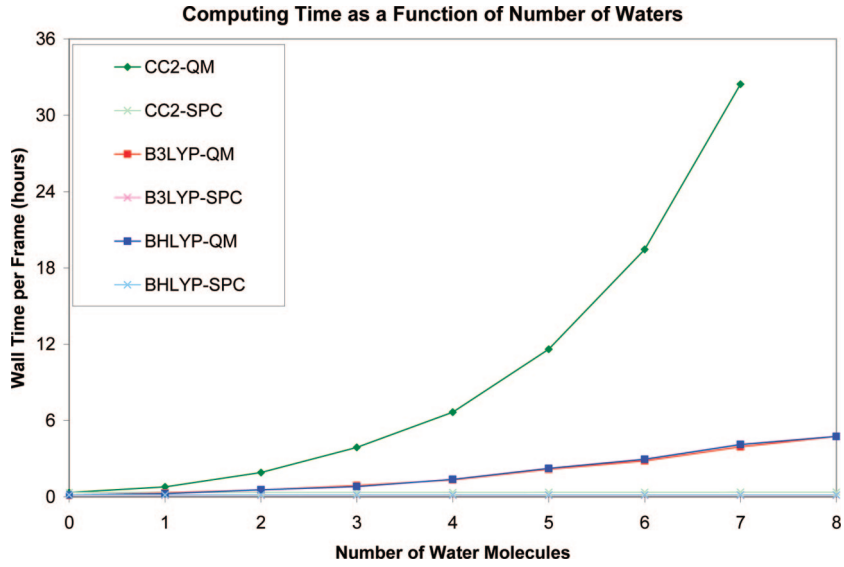



***Figure 6.*** Time required to complete a CD calculation on a single snapshot of a glycine-water system as a function of the number of water molecules. Numbers reported are average values from 128 snapshot simulations. One core of an AMD 2.2 GHz 64-bit dual core Opteron processor was used for each calculation.

a greater extent with the B3LYP functional than with the BHLYP, where it was already much farther away to begin with. Note that the overall excitation energy changes upon full solvation (see Figure 8 below) are roughly comparable for CC2 and B3LYP (slightly larger for B3LYP) and significantly smaller for BHLYP.

**The Correlation between Differing Levels of Theory.** When developing a model chemistry with efficiency in mind, it is helpful to compare the result computed at various methods with those calculated at the most robust level of theory available on a relatively simple system. Here the system of glycine and one water molecule has served as a microcosm of a glycine molecule solvated by multiple water molecules. We have at our disposal to perform multiple calculations on this system at a correlated wave function based level of theory (CC2/aug-cc-pVDZ). However as more waters are added, performing the coupled cluster calculation on an increasingly large system will become impractical and eventually impossible, due to the scaling of this method with respect to system size. This is illustrated quite clearly in Figure 6 where CD calculations (first electronic excitation only) on a system with a glycine and 7 water molecules take over one day *per configuration*, and hundreds of such configurations are required to achieve an averaged result that is a reasonable representation of the dynamic system. With larger systems that include more solvent molecules, modeling with more efficiently scaling methods such as density functional theory or the "zero order scaling" point charge waters becomes a necessity. Employing these point charge waters such calculations take approximately 22 and 8 *minutes* per configuration at the CC2 and DFT methods, respectively, regardless of how many waters are included. As such, it makes sense to take a close look at how closely more efficient model chemistries compare with the coupled cluster model.

In order to do this, we plotted the correlation of the change in the lowest CD excitation (both rotatory strength and wavelength) caused by various types of water molecules versus those changes caused by a QM CC2 water. The QM CC2/aug-cc-pVDZ level of theory is used here as a point of reference since it is the highest level of theory which we determined as practical for the purpose of the present study. The correlation of the change in wavelength is depicted graphically in the top of Figure 7. The correlation of change in rotatory strength is shown in the bottom graph. Each data point corresponds to a distinct snapshot of a glycine in water molecular dynamics simulation; 128 snapshots were considered. For each geometry the glycine molecule was modeled at the QM level, and which ever single water molecule that happened to be closest to the glycine center at that point in time was included as well, either at the QM level or as a set of point charges.

The perturbations to the first excitation of glycine caused by water molecules modeled with all of the less costly methods correlate positively with that caused by a quantum water included at the coupled cluster level. For all of the methods except QM B3LYP, the slope of the correlation in wavelength is less than one, which indicates that the change in excitation wavelength, nearly always a blue shift, caused by the water is smaller in magnitude for the other methods than for full QM CC2. The SPC-CC2 method shows arguably the best correlation with QM-CC2, with a regression line slope of 0.87 and an $R^2$ value over 0.95; the SPC-B3LYP also correlates well, with a slope of 0.87 and an $R^2$ of nearly 0.87. The BHLYP methods, both full QM and SPC, tended to have their CD less affected by water than the other methods, principally since the first excitation with this DFT hybrid is already significantly blue-shifted compared to those calculated with the other methods. Note the correlation that is examined closely in Figure 7 is only that of the water molecule that is closest to the amino acid center in each configuration; for water molecules that are farther away, as was already shown in Figure 5, the correlation between methods appears to improve significantly.

As noted the QM-B3LYP method seems to have the poorest correlation with the QM-CC2 method. It is the only method that shows a change in excitation wavelength that

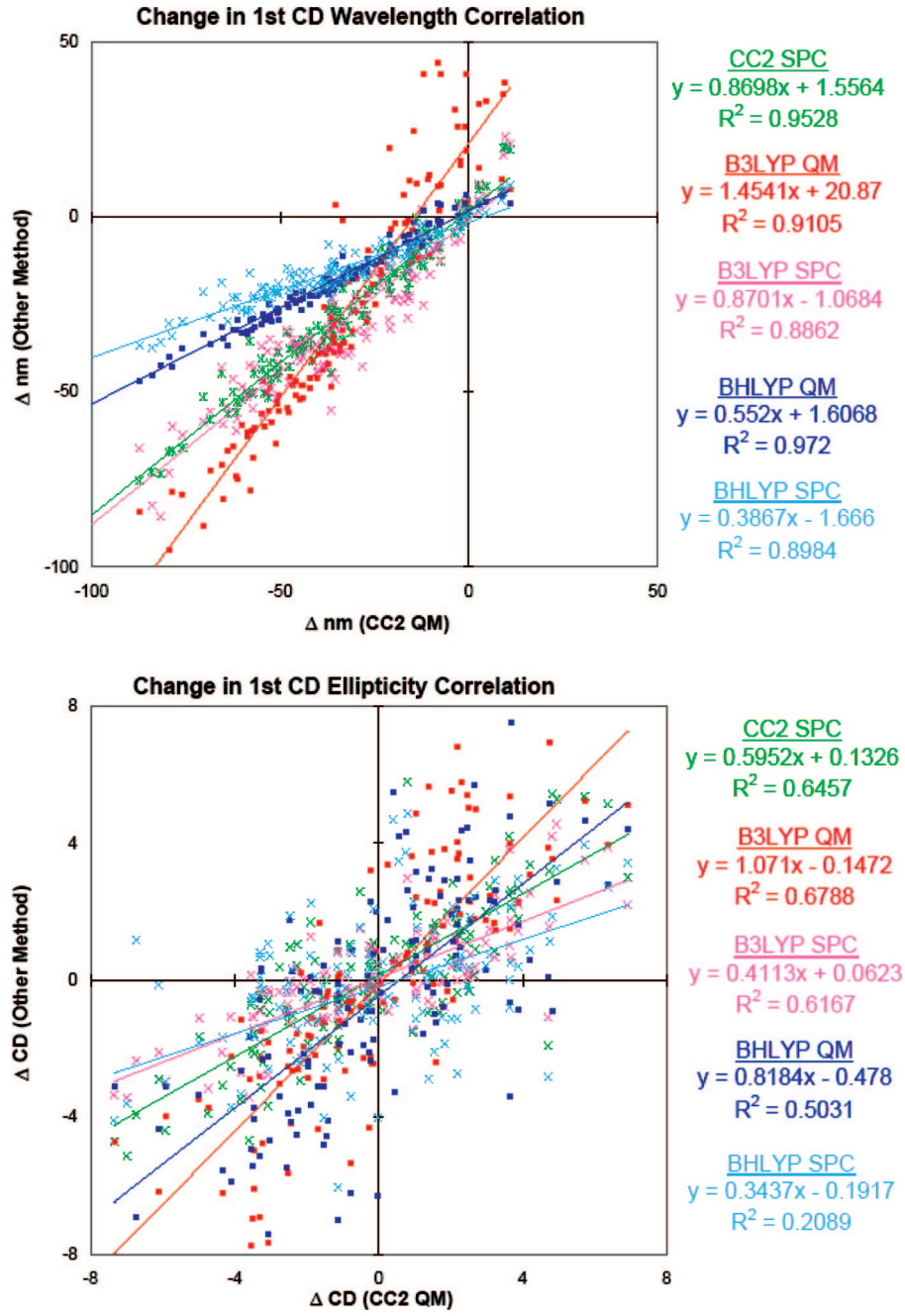



***Figure 7.*** Correlation of wavelength (top) and ellipticity (bottom) of the first excitation caused by the closest water molecule to the glycine solute at various levels of theory with full CC2/aug-cc-pVDZ. 128 configurations were considered.

tends to be greater in magnitude than for QM-CC2. Furthermore, it is the only method that gives a significant number of red shifts to the first excitation for water-glycine configurations where the QM-CC2 method yields blue shifts. Adding a water molecule to a glycine zwitterion should induce a blue shift in its first excitation, due to the stabilization of the glycine electronic state by the solvent. Density functional theory, however, is known to have an issue with producing charge transfer excitations which are often unphysically low in energy. This deficiency, along with its particular consequences in supermolecular solvation simulations such as this one, has been discussed in detail recently by Neugebauer and co-workers.[43] Various methods are under development to compensate for this shortcoming.[44−46] This charge transfer problem can be ameliorated somewhat by using hybrid DFT functionals with a greater portion of exact exchange,[47] such as BHLYP. But with B3LYP, the combination of a QM solute and a QM solvent sets up a scenario where such an charge transfer excitation can take place between solute and solvent or even between two solvent molecules. As indicated by the red data points in the top of Figure 5, it appears to occur frequently enough at this level of theory to be an issue. With point charge waters this is obviously not an issue, since they have no electrons or orbitals to participate in such a nonphysical electron exchange. Thus, with the B3LYP functional the more simplistic point charge waters appear to better model the solvent perturbation of the first electronic excitation of glycine than more costly QM waters do.

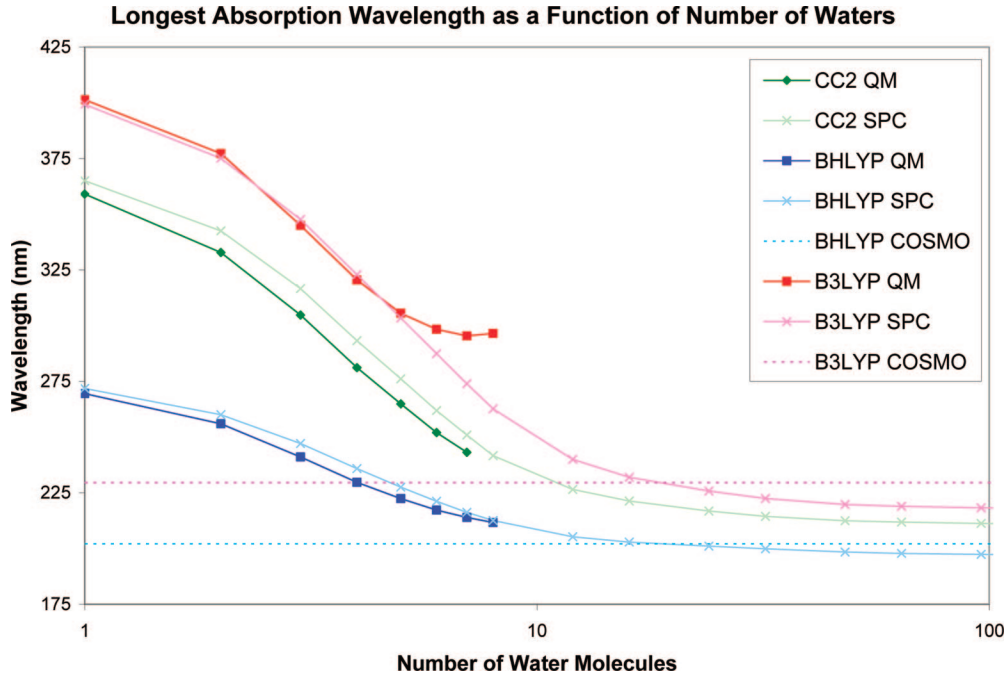



***Figure 8.*** The effects of adding multiple water molecules on the 1st excitation wavelength of the glycine-water system. The results represent the weighted average of 128 configurations of the dynamics simulation of the longest absorption wavelength of various glycine-water systems.

The magnitude of the apparent charge transfer problem is illustrated more dramatically in Figure 8 As one, two, three, and more waters are added to a water glycine system, we expect to see an initial decrease in the wavelength of the lowest electronic transition as the greater number of water molecules stabilize the highly polar glycine zwitterion and increase its HOMO−LUMO gap, with the effect leveling out as the number of water molecules increases. Point charge waters do not cause such unphysical results, and the longest wavelengths computed with these methods converge to slightly shorter wavelengths than those obtained with corresponding continuum based (COSMO) methods. The coupled cluster method does not suffer from the charge transfer problem, so the changes in wavelength at this method is quite similar for QM and SPC waters, regardless of how many are added. BHLYP appears to perform fine as well, though at around eight waters Figure 8 indicates that the charge transfer problem may be beginning to show itself with this QM method as well; using such a hybrid functional with a large portion of exact exchange has indeed compensated for part of the charge transfer excitation problem but has not eliminated it completely. We caution that while the trends shown in Figure 8 are quite consistent with a charge transfer problem, they do not prove that this is the cause or the only cause of the deviations seen in the TDDFT methods. It is also presently unclear if there is any spurious charge transfer present involving water orbitals or if the presence of the water exacerbates an intramolecular charge transfer within the solute.

As for the change in rotatory strength, as with the change in wavelength, there is a positive correlation among all of the solvation methods. At the QM-B3LYP level, the first rotatory strength tends to be larger in magnitude than at QM-CC2, whereas at the QM-BHLYP level it tends to be smaller. This is in keeping with the excitation energies: the CC2 results tend to fall between those obtained with B3LYP and BHLYP. With point charge waters we consistently see perturbations that are too weak compared to QM-CC2. For example, using the SPC charge waters we see a perturbation to the rotatory strength of about 35−60% of that found with full QM CC2 water. This indicates the limits of the point charge water model; it simply cannot reproduce all of the interactions that take place between a water molecule and the solute such as those involving orbital overlap, polarization,[48] and quadrupole and higher order multipole interactions. They can however reproduce the dipole moment of the water molecule, and judging by the slope of the regression lines this seems to be the most important interaction.

Another issue of note is the $R^2$ values of the regression lines. This indication of correlation is rather poor, particularly between the intensities with the BHLYP hybrid and those at QM-CC2. We can see on the bottom graph of Figure 7 many instances where the sign of the CD does not agree between these methods. We have reason to believe that this may be caused by insufficient modeling of electron correlation. In an earlier work, we found that increasing the amount of exact exchange in a hybrid DFT functional eventually resulted in the wrong sign of the CD of the first excitation being modeled.[17] For this modeling of one configuration CD of the alanine zwitterion, this sign change occurred at some point between the BHLYP and HF levels of theory, i.e. where between 50% and 100% exact exchange were used.

**Mixed SPC/QM Solvation.** Thus far, we have only considered the solvation of glycine by point charge or quantum mechanical water molecules but not both at the same time. However a model can be devised in which the water molecules closest to the solute are computed at the QM level, while those farther away are simultaneously

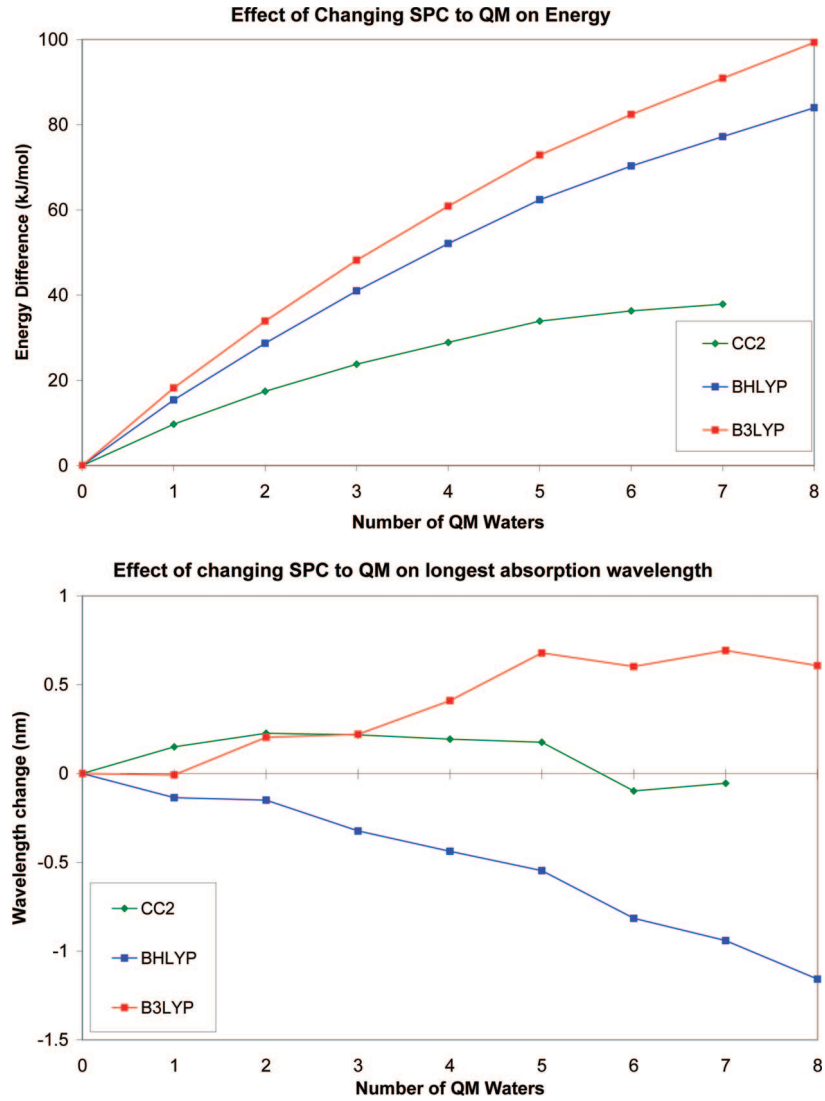



***Figure 9.*** The difference in partial solvation energy and longest absorption wavelength of a glycine molecule solvated by 256 point charge water molecules and one solvated by *n* QM waters and *256-n* point charge waters.

considered as sets of simple point charges. To evaluate the merits of such methods, we performed calculations of a system in which glycine is solvated by 256 point charge water molecules. These results were compared to computations in which glycine is solvated by *n* QM waters, where *n* = 1 to 8, and *256-n* SPC water molecules. Waters were designated as QM based on their proximity to the solute center, with the closest being the first considered as QM, the next closest the second, and so on. The results are depicted in Figure 9 based on an average of 128 MD configurations each.

The discrepancies in close-distance solvation energies between point charge and quantum waters that was first noted in Figure 3 is even more apparent in the top of Figure 9. The water molecules considered in Figure 9 all reside in the first solvation shell, where they are apt to be in close contact with the solute, so we expect to see the most pronounced differences here. For the waters in closest contact with the glycine we see that the point charge waters result in a greater stabilization than do the QM waters, which is not unexpected since point charge waters do not take into account some types of interactions, such as steric repulsion, which are more prominent at short distances. The point charge waters are far more similar to their corresponding QM waters with the CC2 method than with the DFT, though such differences are still significant. However we can see at all levels of theory the significance of using a QM water as opposed to a point charge water diminishes as the waters in question get progressively farther from the solute center.

The effect of using QM as opposed to SPC water molecules for the innermost solvation is far less noticeable on the first excitation wavelength. As is shown in the bottom of Figure 9, switching from point charge to QM water molecules for even the innermost 8 water molecules of the 256 molecule solvation sphere results in a maximum change of around 1 nanometer in wavelength. At the CC2 level of theory, which should be the most reliable, this difference between QM and SPC waters is quite negligible, ~0.2 nm. This indicates that even for the closest held waters of the solvation sphere a simple point charge model

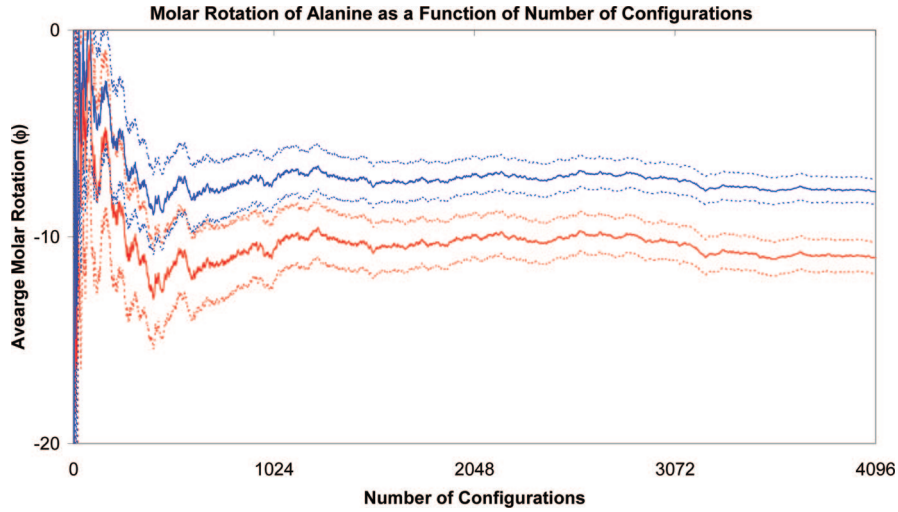



***Figure 10.*** The convergence of the molar rotation of alanine as the number of averaged configurations increases. The BHLYP/SPC (blue) and B3LYP/SPC (red) methods were used. The running averages are represented by solid lines, while the error bars, at one standard deviation, are represented by dashed lines.

***Table 1.*** Molar Rotation ([$\phi$], deg·cm$^2$/dmol) of Glycine and Alanine with Various Methods[a]

| | | | static COSMO | | dynamic COSMO | | dynamic SPC | |
|---|---|---|---|---|---|---|---|---|
| molecule | | expt | B3LYP | BHLYP | B3LYP | BHLYP | B3LYP | BHLYP |
| glycine | $\phi$ | 0 | 0.2 | 0.1 | −0.9 | −0.5 | −1.0 | −0.8 |
| | ± | | | | 1.4 | 0.9 | 1.1 | 0.8 |
| alanine | $\phi$ | 1.6 | 5.4 | −1.5 | −16.7 | −10.9 | −11.0 | −7.8 |
| | ± | | | | 0.8 | 0.6 | 0.8 | 0.6 |

[a] Experimental value for alanine is from ref 50. Calculations with static molecules were done by the same method used in our prior work.[16] Error bars (±) represent one standard deviation of statistical error from the dynamics simulation.

is quite capable of modeling the solvent effects on this electronic excitation.

**Glycine and Alanine: Comparing Molar Rotations with Experiment.** This work so far has focused on the chiroptical response properties of glycine, a molecule that may serve as a means for calibrating our method but which is itself achiral. This serves the purpose of a "blank" for our molecular dynamics method of modeling molar rotation. The molar rotation of glycine must average to zero, and so whatever residual molar rotation that does not average out after computations on multiple snapshots of the glycine in water system will indicate some idea of how much statistical error is inherit in the method. It seems prudent here to also take this method and extend its application to a similar chiral amino acid, one whose chiroptical response properties ought not to average to zero over time and whose chirality, as exhibited in its molar rotation, can be modeled and compared to that of experiment. Alanine, the smallest *chiral* amino acid, serves as a natural target for such an investigation.

In this section 40960 ps molecular dynamics runs of glycine and alanine were performed under the same conditions as the glycine simulations in the foregoing sections. A total of 4096 equally spaced configurations were taken for subsequent molar rotation calculations. Both the B3LYP and BHLYP hybrid DFT methods were employed, using either our simple point charge waters or the COMSO continuum model that we used in prior works.[16,17,19] For each selected configuration along the molecular dynamics trajectories a molar rotation was computed, and that rotation was appended to a running average of molar rotations. Two examples, alanine in SPC waters with the BHLYP and B3LYP functionals, are shown in Figure 10. Convergence criteria are based upon the principles of signal averaging, under which the standard deviation of a data set should drop as $n^{1/2}$ per number of points averaged, n, if the points are uncorrelated.[49] We confirmed that for these systems this is the case when we allowed 10 ps to elapse between snapshots of the molecular dynamics run.

These results are compared with those obtained using static alanine molecules and the COSMO model as well as with those from experiment. These data and data from calculations on glycine are summarized in Table 1.

The data in Table 1 show that the molecular dynamics/point charge solvation method produces results comparable to earlier results with frozen solute molecules and a continuum solvent. All results are reasonably close to experiment and well within the margin of error typical of TDDFT based calculations.[51] Error bars tended to be larger with the COSMO solvent model than with the discrete SPC waters and higher for the B3LYP hybrid than for the BHLYP. This is due to the relatively lower excitation energies obtained with COSMO and with B3LYP, which result in larger magnitudes of the computed molar rotations.

In addition to comparing our results to experiment, we should also mention the results of D'Abramo et al., who

performed dynamics based computations of alanine in water using TDDFT and a "perturbed matrix method" of solvation.[14] Using this method they obtained a specific rotation of about +60 deg·cm$^3$/(g·dm) for alanine, which corresponds to a molar rotation of +53 deg·cm$^2$/dmol. It would appear that our method is significantly closer to experiment. However, their rotatory dispersion calculations were performed with a truncated sum-overstates method. In contrast, the linear response method used here does not possess such truncation errors.[52] As such, a direct comparison between results with our point charge solvation model and their perturbed matrix model is not currently possible, though the disagreement is likely due to the truncation error from the sum over states calculation of alanine.[16]

## Conclusions

Simple point charge (SPC) water molecules have been shown to be a computationally efficient alternative to using quantum mechanical waters in modeling the solvent effect on a solute's chiroptical responses. The near zero order scaling of point charge waters allows hundreds of explicit water molecules to be considered at negligible computational cost. The effect that SPC waters have on the computed chiroptical properties of a solvated glycine molecule are comparable to those obtained with the much more expensive CC2 method. When density functional theory is employed, point charge waters may prove superior to explicit QM waters in simulating solvent effects on such response properties, since the point charge model does not exacerbate the problem of DFT with spurious charge transfer excitations. Calculations on the chiral amino acid alanine with the SPC/DFT method yield results that are in reasonably good agreement with experiment and marginally better than those with the same geometries and the COSMO/DFT method. Further benchmarking of the point charge/DFT protocol is presently underway.

**Acknowledgment.** The authors would like to acknowledge support from the Center for Computational Research (CCR) at the University at Buffalo for computational resources. M.D.K. wishes to thank Marcel Swart, Marek Freindorf, and Mark Rudolph for helpful discussions regarding molecular dynamics programs. J.A. is grateful for financial support from the CAREER program of the National Science Foundation (CHE-0447321).

### References

(1) Sahu, P. K.; Lee, S. L. *J. Mol. Model.* **2008**, *14*, 385.

(2) Aikens, C. M.; Gordon, M. S. *J. Am. Chem. Soc.* **2006**, *128*, 12835.

(3) Takahashi, H.; Kawashima, Y.; Nitta, T.; Matubayasi, N. *J. Chem. Phys.* **2005**, 123.

(4) Ramaekers, R.; Pajak, J.; Lambie, B.; Maes, G. *J. Chem. Phys.* **2004**, *120*, 4182.

(5) Balta, B.; Aviyente, V. *J. Comput. Chem.* **2004**, *25*, 690.

(6) Balta, B.; Aviyente, V. *J. Comput. Chem.* **2003**, *24*, 1789.

(7) Cui, Q. *J. Chem. Phys.* **2002**, *117*, 4720.

(8) Gontrani, L.; Mennucci, B.; Tomasi, J. *Theochem-J. Mol. Struct.* **2000**, *500*, 113.

(9) Kassab, E.; Langlet, J.; Evleth, E.; Akacem, Y. *Theochem-J. Mol. Struct.* **2000**, *531*, 267.

(10) Tortonda, F. R.; Pascual-Ahuir, J. L.; Silla, E.; Tunon, I.; Ramirez, F. J. *J. Chem. Phys.* **1998**, *109*, 592.

(11) Tortonda, F. R.; PascualAhuir, J. L.; Silla, E.; Tunon, I. *Chem. Phys. Lett.* **1996**, *260*, 21.

(12) Jensen, J. H.; Gordon, M. S. *J. Am. Chem. Soc.* **1995**, *117*, 8159.

(13) Kikuchi, O.; Watanabe, T.; Ogawa, Y.; Takase, H.; Takahashi, O. *J. Phys. Org. Chem.* **1997**, *10*, 145.

(14) D'Abramo, M.; Aschi, M.; Di Nola, A.; Amadei, A. *Chem. Phys. Lett.* **2005**, *402*, 559.

(15) Sagarik, K.; Dokmaisrijan, S. *Theochem-J. Mol. Struct.* **2005**, *718*, 31.

(16) Kundrat, M. D.; Autschbach, J. *J. Am. Chem. Soc.* **2008**, *130*, 4404.

(17) Kundrat, M. D.; Autschbach, J. *J. Phys. Chem. A* **2006**, *110*, 4115.

(18) Krykunov, M.; Kundrat, M. D.; Autschbach, J. *J. Chem. Phys.* **2006**, *125*, 13.

(19) Kundrat, M. D.; Autschbach, J. *J. Phys. Chem. A* **2006**, *110*, 12908.

(20) Berova, N.; Nakanishi, K.; Woody, R. W. *Circular Dichroism: Principles and Applications*, 2nd ed.; John Wiley & Sons, Inc.: New York, 2000.

(21) Barron, L. D. *Molecular Light Scattering and Optical Activity*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 2004.

(22) Mennucci, B.; Tomasi, J.; Cammi, R.; Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Gabriel, S.; Stephens, P. J. *J. Phys. Chem. A* **2002**, *106*, 6102.

(23) Cappelli, C.; Monti, S.; Rizzo, A *Int. J. Quantum Chem.* **2005**, *104*, 744.

(24) Pecul, M.; Larnparska, E.; Cappelli, C.; Frediani, L.; Ruud, K. *J. Phys. Chem. A* **2006**, *110*, 2807.

(25) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(26) Mukhopadhyay, P.; Zuber, G.; Goldsmith, M. R.; Wipf, P.; Beratan, D. N. *ChemPhysChem* **2006**, *7*, 2483.

(27) Mukhopadhyay, P.; Zuber, G.; Wipf, P.; Beratan, D. N. *Angew. Chem. Int. Ed.* **2007**, *46*, 6450.

(28) Jensen, L.; Swart, M.; Van Duijnen, P. T.; Autschbach, J. *Int. J. Quantum Chem.* **2006**, *106*, 2479.

(29) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

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

(31) Christiansen, O.; Koch, H.; Jorgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.

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

(33) Schafer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. *PCCP Phys. Chem. Chem. Phys.* **2000**, *2*, 2187.

(34) Pedersen, T. B.; Koch, H.; Boman, L.; de Meras, A. *Chem. Phys. Lett.* **2004**, *393*, 319.

(35) Crawford, T. D.; Tam, M. C.; Abrams, M. L. *J. Phys. Chem. A* **2007**, *111*, 12057.

(36) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.

(37) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

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

(39) Destro, R.; Roversi, P.; Barzaghi, M.; Marsh, R. E. *J. Phys. Chem. A* **2000**, *104*, 1047.

(40) Isborn, C.; Claborn, K.; Kahr, B. *J. Phys. Chem. A* **2007**, *111*, 7800.

(41) Legrand, M.; Viennet, R. *Bull. Soc. Chim.* **1965**, 679.

(42) Nishino, H.; Kosaka, A.; Hembury, G. A.; Matsushima, K.; Inoue, Y. *J. Chem. Soc., Perkin Trans. 2* **2002**, 582.

(43) Neugebauer, J.; Louwerse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 13.

(44) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.

(45) Neugebauer, J.; Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2006**, *124*, 11.

(46) Lange, A.; Herbert, J. M. *J. Chem. Theory Comput.* **2007**, *3*, 1680.

(47) Bernasconi, L.; Sprik, M.; Hutter, J. *Chem. Phys. Lett.* **2004**, *394*, 141.

(48) Osted, A.; Kongsted, J.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2004**, *108*, 8646.

(49) Coutinho, K.; De Oliveira, M. J.; Canuto, S *Int. J. Quantum Chem.* **1998**, *66*, 249.

(50) Greenstein, J. P.; Winitz, M. *Chemistry of the Amino Acids*; John Wiley & Sons: New York, 1961.

(51) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J. *J. Phys. Chem. A* **2001**, *105*, 5356.

(52) Polavarapu, P. L. *Chirality* **2006**, *18*, 348.

CT8002767

JCTC Journal of Chemical Theory and Computation

# Multiscale Modeling of the Atomic Layer Deposition of $HfO_2$ Thin Film Grown on Silicon: How to Deal with a Kinetic Monte Carlo Procedure

A. Dkhissi,* A. Estève, C. Mastail, S. Olivier, G. Mazaleyrat, L. Jeloaica, and M. Djafari Rouhani

*Laboratoire d'Analyse et d'Architecture des Systèmes-CNRS, University of Toulouse, 7 avenue du Colonel Roche, 31077 Toulouse, France*

Received April 09, 2008

**Abstract:** An original integrated approach developed within a multiscale strategy, which combines first-principles quantum simulations and kinetic Monte Carlo (KMC), is presented to investigate the atomic layer deposition (ALD) of $HfO_2$ on Si(100) surface. Density functional theory within the hybrid functional is used to determine the detailed physicochemical mechanisms and associated energetics of the two half cycles taking place during the initial stage of film growth. A kinetic Monte Carlo model is then proposed that deals with the stochastic nature of the calculated DFT mechanisms and barriers. Beyond the chemical information emanating from DFT calculations, the lattice-based KMC approach requires preliminary physical considerations issued from the crystal structures that the system is intended to adopt. This is especially critical in the case of heterogeneous systems like oxides deposited on silicon. We also describe (i) how atomistic configuration changes are performed as a result of local events consisting in elementary reaction mechanisms occurring on specific lattice sites, (ii) the temporal dynamics, governed by transition probabilities, calculated for every event from DFT activation barriers, and (iii) the relation of KMC with the ALD experimental procedure. Some preliminary validation results of the whole multiscale strategy are given for illustration and pertinence with regard of the technological main issues.

## 1. Introduction

In the electronic field, $SiO_2$ has been the gate dielectric of choice for MOS devices for several decades because of its ability to grow on silicon, its thermal stability, and its low level of defect density at the interface, resulting in excellent electrical properties of the devices. Unfortunately, the reduction of the $SiO_2$ gate oxide thickness in agreement with Moore's law has led to unacceptable tunnelling and leakage current levels.[1] So, the conventional $SiO_2$ gate reaches its physical and electrical limitations. An intense effort to find a replacement for $SiO_2$ as the gate dielectric for future MOS electronics has been under way for several years. For this purpose, the importance of high-*k* gate dielectrics has been well demonstrated.[2] Indeed, in addition to showing a high dielectric constant, a potential high-*k* replacement must satisfy many stringent requirements, such as the high quality of the interface, a desired band alignment with silicon, and a large electronic gap. At present, the leading materials are $Al_2O_3$, $HfO_2$, and $ZrO_2$ oxides, considered as the first high-*k* generation candidates to meet these criteria and replace $SiO_2$. For MOS applications, $HfO_2$ is attractive because it exhibits a bulk permittivity of almost 25, a wide band gap (5.68 eV), and a good thermodynamic stability in contact with silicon.[3] However, the growth of a $SiO_2$ interfacial layer between $HfO_2$ and the silicon substrate leads to an increase of the experimental EOT (equivalent oxide thickness) of the gate stack.

Among various methods for growing high-*k* dielectric films, atomic layer deposition (ALD) show a unique ability to deposit ultra thin films with excellent conformity and uniformity over large areas.[4] ALD, which is a vapor

* To whom correspondence should be addressed. E-mail: adkissi@laas.fr.

10.1021/ct8001249 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

deposition technique, is based on the cycling of self-terminating surface reactions. Indeed, each precursor is pulsed into the reaction chamber alternately, and the reaction between the incoming precursors and surface species is self-terminating. Thus atomic-level control of film growth is supposed to be achieved. Experimentally, ALD has been actively investigated for deposition of $HfO_2$[5−21] for which $HfCl_4$ and $H_2O$ are often used as precursors.[22−24] In practice, the quantity of metal atoms deposited per cycle depends on the temperature, on the chemical nature of the precursors used, and related reactive sites on the surface. For instance, Ritala et al. found that only a submonolayer of the $HfO_2$ film is deposited during each cycle.[22] Therefore, a detailed understanding of the basic mechanisms that take place during each cycle is required if one wants to optimize the deposition and to reach full monolayer coverage at each cycle. This is especially true for the few first deposited layers whose quality, in terms of interface defects, is of major importance for subsequent electrical properties. In recent studies, some groups have investigated specific problems of ALD of hafnia films, their structural, optical, and electrical characteristics.[25−28] They show that diffusion and solid-state reactions at the substrate−film interface may influence the film material and create or modify an interface layer between the film and substrate.[29] On silicon substrates, these reactions could result in undesirable formation of defects, silicate or silicide layers, which may further significantly deteriorate the properties of the dielectric layer.[29−32] Others issues, such as contamination effects resulting from the uncontrolled precursor decomposition, are observed. In particular, chlorine contamination has been asserted and well documented on films grown with $HfCl_4$ precursors.[20,33−35] It is of great importance for optimal future experimental setups that a precise atomic level description of the basic reaction mechanisms responsible for the overall process of high-*k* film growth and also that their relation with the thermodynamic parameters be established. In this respect, the decomposition of $HfCl_4$ on the substrate and further hydrolysis of the resulting surface complex, that is, $SiO_2$−O−$HfCl_3$ should be thoroughly investigated. Meanwhile, today's state-of-the-art theoretical methods can play a decisive role in reduction of the expensive experimental efforts needed for screening large numbers of candidate materials and associated process parameters. Theoretical approaches can speed up the selection of suitable gate dielectrics and growth methods by limiting the experimental input needed. For these reasons, a detailed understanding of the basic mechanisms that take place during each cycle of ALD has been the subject of intensive research effort by several theoretical groups using first principles, especially via density functional theory.[35−51] In our group, we have considered the ab initio study of the initial stage of the ALD as being essential for the complete understanding and the control of the growth of the high-*k* ($Al_2O_3$, $ZrO_2$, and $HfO_2$) films for gate dielectrics applications.[49−51] Our final aim is to achieve the ambitious task of elaborating a new generation of tools dedicated to atomic scale simulation of technological process, for an optimization of experimental setups. Here, our strategy is to combine the ab initio level of calculations with kinetic Monte-Carlo techniques into a multiscale approach, which will incorporate enough reaction mechanisms to be ready for being tested against experimental setups. Rigorously, molecular dynamics should be a suitable tool to address atomic scale process simulation. However, microelectronic processes deal with mesoscopic structures, several millions of atoms, and cover timescales beyond seconds or minutes. Molecular dynamics is then limited for two major reasons: (i) interatomic potentials have to be generated, which is a particularly difficult task in the frame of heterostructures and interfaces, and (ii) timescales are limited to nanoseconds of experimental duration. Thus kinetic Monte Carlo (KMC) is an alternative procedure that is of great help in the modeling of microelectronic processes: a lattice-based model makes it possible to avoid the detailed description of continuous atomic trajectories, and the kinetics of the sequence of events is explicitly treated and can meet typical process durations. While traditional thermodynamic and kinetic models deal with average physical quantities, kinetic Monte Carlo is able to consider a wide range of possible configurations at the atomic scale and to choose only one random path out of all possible ones. This corresponds to an actual (ALD) experiment. The path is determined according to random numbers sampled according to transition probabilities between configurations. Obviously, the transition probabilities depend also on the local configuration, activation barriers and on the film deposition conditions, such as pressure and temperature. Moreover the role of each mechanistic step on an ensemble of interacting species makes it feasible to uncover process-dependent types of growth, kinetics, and their associated atomic arrangements. The present paper is dedicated to the description of the preliminary version of our KMC tool enabling the treatment of $HfO_2$ ALD from $HfCl_4$ and $H_2O$ precursors at the atomic scale. In the following, we will focus our attention on the basic ingredients needed to develop what we call a lattice-based KMC model. We will define a lattice framework able to represent the transition from the silicon diamond structure to the $HfO_2$ crystalline structure. Then we introduce the concept of basic atomistic mechanisms and show how they can be derived from ab initio calculations. We will finally present the temporal dynamics rules of our model before showing validation and applications.

## 2. Theoretical Approaches

**2.1. Ab Initio DFT.** The calculations are performed within the framework of Kohn−Sham density functional theory. Specifically, we used the hybrid functional B3-LYP, which combines Becke's three-parameter exchange functional (B3)[52] with Lee−Yang−Parr gradient corrected correlation functional (LYP).[53] All the atoms are described by a triple-$\zeta$ valence plus polarization (TZVP) functions. For Hf atoms, we used the ecp-60-mwb energy-consistent pseudopotentials given by the Stuttgart−Dresden−Bonn (SDB) group.[54−56]

The surface is built according to the following procedure. It is modeled by a cluster that includes a dimer unit of the $2 \times 1$ reconstructed Si(100) surface having each of its Si dangling bonds passivated by an H atom. Only the first Si

layer atoms are left free during the relaxation. The second-, third-, and fourth-layer atoms, including the H terminations, are held fixed in their bulk tetrahedral positions. This type of constraint will avoid unrealistic relaxations of the surface model during the optimization procedure. The unconstrained part of the cluster was then oxidized by five oxygen atoms placed in the Si−Si bonds closest to the surface. Therefore, subsequent relaxations were performed after each newly arrived O atom was embedded. The result is 1-dimer peroxide-like[57,58] model for $Si(100)/SiO_2$ $2 \times 1$ surface. Of the two remaining dangling bonds on the dimer, one is hydroxylated and the other hydrogenated. One then obtains a 28-atom optimized structure with a final stoichiometry of $Si_9O_6H_{13}$. The calculations are performed with Turbomole5.5[59] and Gaussian03.[60]

**2.2. Kinetic Monte Carlo (KMC).** KMC is a stochastic-based model aimed at simulation of film growth at the atomic scale, the final objective being to furnish alternative/new tools for replacement of the macroscopic conventional TCAD tools used for years by engineers in microelectronics. Indeed, from the engineering point of view, the main interest is in the microstructure that is produced under specific processing conditions. Such a model performs virtually the explicit experimental processing procedure event after event, at the atomic scale. This is operated through probabilistic rules that make the overall simulation comparable with an actual experiment.[61,62] It involves millions of atoms with a time duration longer than the second. Unfortunately, the most predictive models, that is, the quantum-based models are not tractable at this scale. Two options can be further considered: molecular dynamics (MD) and kinetic Monte Carlo (KMC). MD solves the Newton equation of motion for each atom in the system. Trajectories are continuous and necessitate very short integration time steps that make the simulation cumbersome for durations beyond the nanosecond. Another drawback of the MD method is the general lack of adequate interatomic potential for highly disordered and heterogeneous systems. This is particularly true for microelectronics semiconductor/oxide interfaces such as silicon/high-*k* gate oxides. KMC appears to be a simplification compared with MD: MD continuous trajectories are replaced with discrete atomic jumps. A prior knowledge of these hoppings is needed and must be listed from any other source: quantum-based models or experimental characterization. Thus, provided the characteristics of reaction pathways are known, KMC methods allow the simulation of more ambitious systems in terms of size and duration of the simulated experiment, meeting the requirements of the next generation of processes in microelectronics. Also, the lattice-based KMC, because of its induced structural simplification, makes it easier to process the data concerning the atomic arrangement as a function of the implemented basic mechanisms. This contrasts with the difficulty in dealing with structural aspects in MD, where complex statistical methods have to be employed.

Basic ingredients needed to develop what we call a lattice-based KMC model are detailed as follows: (i) we first define a lattice framework able to operate the desired transition between the silicon crystal and the oxide structure; (ii) we then characterize each configuration with the aim of associating each lattice site to the chemical nature of the species occupying the site; (iii) we introduce the concept of events that must describe correctly the chemistry of the basic mechanisms; (iv) we finally indicate how to deal with the time evolution in the KMC, namely, the temporal dynamics of the KMC.

*2.2.1. One Crystal Lattice Model. 2.2.1.1. Lattice Description.* In contrast to molecular dynamics, where atoms are moving continuously in space, the lattice-based procedure developed in this KMC software allows a very efficient treatment of atomic displacements: the atoms discretely move from predefined sites to other sites according to transition probabilities. Obviously, this schematic picture is motivated by physicochemical considerations through the knowledge of the basic mechanisms and the investigation of lattice structure. This last point becomes crucial in the case of heterogeneous systems, where at least two different crystallographic structures are being considered. This is the case in our study, where metallic oxides are grown on an ultrathin silicon oxide, itself grown or deposited on a silicon substrate. First, a reliable and systematic way of representing the atom locations is needed. Then, the implementation of configurations makes the connection between a location and its chemical nature: for instance, a site may be unoccupied or occupied by Si, Hf, etc.We point out that this lattice picture gives the possibility of making schematic and comprehensive graphical views of the system and therefore of performing a simple analysis of the results that does not require complex post-processing of the data. In this description, where an atom is represented in its network site, its real location is implicitly somewhere around this particular site.

*2.1.1.2. Modeling of the Atomic Configuration.* We now describe the construction of a model in charge of representing the atomic configuration of a $Si/SiO_2/HfO_2$ system of sites. Beyond the crystalline aspects, we will introduce tools able to take into account the molecular states: precursors, substituents, contaminants, etc. The management of the neighbors will also be investigated. Attention will finally be given to the modeling of the substrate and its connection with the introduced crystalline model. Overall crystallinity is defined when Hf and oxygen atoms are located in the predefine lattice sites. In this case, Hf may be fully or under-coordinated depending on the occupancy of the oxygen lattice sites around. When densification occurs, rearrangement of oxygen atoms should induce an increase of the coordination number.

*a. Crystallographic Considerations.* A preliminary crystallographic study is essential in close connection with the construction of the atomic configuration. In our case, the system is heterogeneous: "high-*k*" oxide on silicon with an ultrathin silica interface. Moreover, the basic species are molecular precursors whose mechanisms of decomposition prove to be complex and poorly understood. It is thus necessary to develop a system of location able to represent simultaneously various crystalline structures containing various elements. Moreover, further refinements will be necessary to take account of several subtleties, such as molecular states, substituents, and contaminants, in opposition to crystalline states. First of all, it is advisable to know the crystallography of the $HfO_2$ oxide. Hafnia exists under

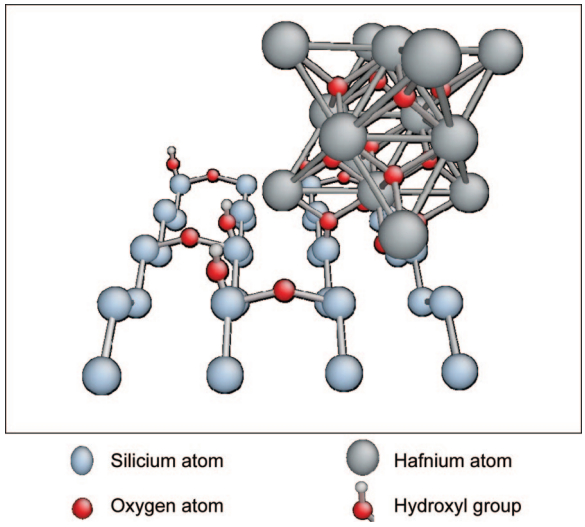



***Figure 1.*** Cubic cell of hafnia deposited on the $Si/SiO_2$ substrate functionalized with hydroxyl groups.

different crystallographic structures. At low temperature, $HfO_2$ has a monoclinic phase. When the temperature is increased, it transforms into the tetragonal phase and then into the cubic phase. Experimental evidence of crystalline grains for thick oxide or post-annealed films has been established. However, the structure close to interface, that is, the oxide structure near the silicon/silicon dioxide, is not known. Therefore, to correctly model this interface, one has to make a coincidence lattice site study. For us, it then appears judicious to consider the (100) face of the cubic phase of hafnia: it matches efficiently the (100) silicon surface structure, is geometrically easier to describe, and does not contradict or hinder the experimental results. Thus, the overall idea consists of building the model of location via a single cubic structure, keeping in mind that the real positions of atoms are "somewhere around" these arbitrary crystalline positions. The substrate will be a Si(100) surface functionalized by hydroxyl groups, which are known to be the active surface sites with regard to the precursor gas phase. [63]

*b. Two-Dimensional Basic Cell.* To characterize the sites in our KMC, we have to define a two-dimensional cell where there should not be any ambiguity between atom species and locations. For that, we superimpose the conventional cubic cell of oxide and silicon-based $SiO_2$ (100) surface and try to find an agreement between their crystallographic parameters. The silicon surface structure is taken as the crystal reference, an ultrathin silicon dioxide being able, in this view, to accommodate perfectly with this structure. We obtain the hypothetical configuration represented in Figure 1. This resulting picture leads to the best agreement between the different cells. The silicon substrate is represented by two atomic layers to know the orientation of the bonds in the higher layers. In Figure 1, siloxane bridges and hydroxyl functions are represented just as an indication. This surface representation is very schematic: in particular, we do not show the dimer formation inducing Si−O−Si species, although they are taken into account in the calculations. Silicon atoms are schematically located at the nodes of the diamond lattice, without taking account of displacements usually observed. According to the crystallographic tables,[64]

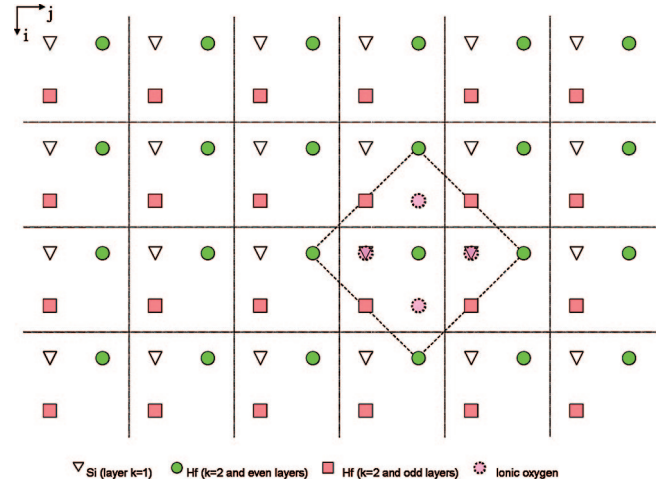



***Figure 2.*** Top view of ideal configuration. Several $HfO_2$ cubes are deposited on the substrate with an ideal agreement of cell. A two-dimensional cell is identified where there should not be any ambiguity between atom species and locations.

the distance separating two Si surface atoms is 3.84 Å. If one then considers that the cubic cell parameter of hafnia oxide is 5.04 Å, the distance separating two neighboring Hf atom is 3.57 Å. It thus appears possible to accommodate the two structures, thanks to the coincidence between Si(100) and Hf(100), with a tension of the cubic hafnia that can be assimilated to an intrinsic strain at the interface level of the. In more detail, how do Hf positions compare to those of Si? The answer that we formulated is illustrated in Figure 1.

Let us imagine now that we deposit several $HfO_2$ cubes in tension, as seen previously, on the substrate, with an ideal agreement of the cells. We see that a tilt of the cell is required to obtain a complete matching of the atoms. It is then possible to completely cover the surface without defect. It is, to some extent, what we would like to obtain in real experiments, the defects resulting from the kinetic process and the stochastic nature of the growth. Figure 2 is a top view of this ideal configuration where the various species are represented by symbols for a better legibility. We thus have a (100) surface of silicon atoms, covered by single-crystal $HfO_2$ in its cubic phase. The square base of a conventional cubic cell is represented in dotted line. Five hafnium atoms are located in the second layer ($k = 2$): the four corners of the square and the central atom (gray round symbols). On the monolayer above ($k = 3$), we find four hafnium atoms corresponding to the centers of the four side faces (gray square symbols). The first four oxygen atoms (gray round symbols on dotted line) are situated between these two layers ($k = 2 + {}^1/_2$). By addition of two new layers, one of oxygen atoms (gray round symbols on the dotted line at $k = 3 + {}^1/_2$) and one of hafnium atoms (gray round symbols at $k = 4$), one obtains the totality of the conventional cube. This cube is obviously not the primitive cell, minimal for the construction of a crystal. The primitive cell, which we consider to be the unit cell, is represented in thick dotted lines and is found by observation of Figure 2 globally. According to altitude, or more exactly to the layer index $k$, one meets various species, placed at various positions of this cell. For the $k = 1$ layer, we have a silicon placed on a

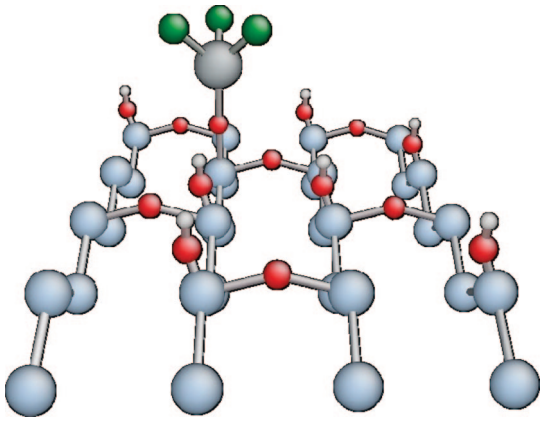


***Figure 3.*** Noncrystalline −O−$HfCl_3$ groups chemisorbed on the hydroxylated $Si/SiO_2$. This is the KMC model structure

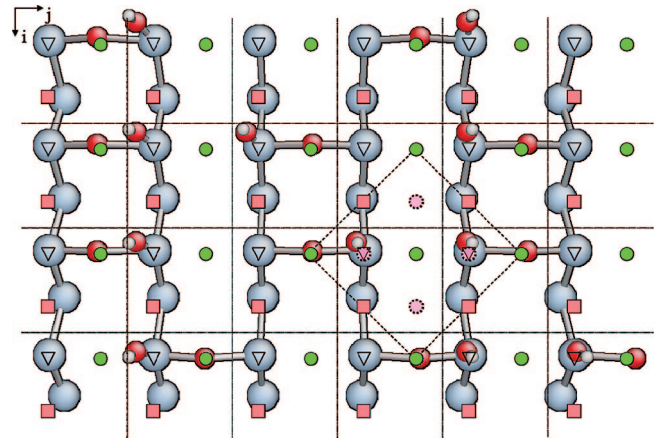



***Figure 4.*** Connection between the atomic model of configuration and the substrate.

triangle. For the even layers, a hafnium will be placed on a gray round symbol, and for the odd layers (except $k = 1$), hafnium will be located on a dark gray square. Between two layers of hafnium, we find two oxygen atoms placed on the gray round dotted lines (for clarity, they are represented only in the conventional cell). This primitive cell thus makes it possible to represent in an exhaustive way an $HfO_2$ crystal, in its cubic phase, on a (100) silicon surface. It contains preset sites that are able to accommodate the various atoms that compose the system, while placing them at the adequate position.

*c. Atomic Configurations.* Through the lattice description, we built a grid to help us to locate the atoms. Now we must formalize the possible chemical nature of species occupying these sites. Thus, the configuration, that is, the site occupation nature, should characterize the local atomic arrangement within the site. We introduce a data-processing structure that will take account not only the crystalline configuration but also the chemical functions, the various contaminants, and the molecular groupings with their substituent. Each element of grid will thus contain several fields of occupation. For the metal element, it is provided in molecular form in the precursor, is then incorporated as an adsorbed molecule, and finally, is stabilized in the crystalline state. Therefore, the field of metal occupation can be empty (value 0), be occupied by a molecular hafnium (value 1), or occupied by a crystalline hafnium (value 2). A crystalline state corresponds to the case where a hafnium is located somewhere around a preset site, while being based on the cubic cell of dense oxide. On the other hand, in its molecular state, the Hf atom is attached to the substrate by only one connection, that is, a treelike structure, after physisorption or chemisorption of the precursor molecule. It is advisable to distinguish these two states because their reactivities will be, without any doubt, different because one Hf−OH or Hf−Cl dangling bond transforms into Hf−O−Si−substrate (Sub) tightly bound to the substrate. As an example, Figure 3 presents one such molecule, a chemically absorbed grouping, having three Cl substituents attached. The substituents have their own fields of occupation: one for Cl and one for OH. The values of these two fields can reach 4; their sum has to remain lower than 4 to respect the coordination number of hafnium atoms in their covalent environment. For the crystalline oxygen atoms considered in the construction of the elementary cell, two fields of $O_a$ and $O_b$ sites are identified. They can be empty or occupied. Let us recall that these oxygen atoms are of crystalline type and do not have to be confused with those of hydroxyls or the groupings, O−Hf, which are considered as molecular.

These oxygen crystalline sites can be occupied after what we will call "densification" mechanisms, described later in Section 3.1.3.The reaction mechanisms presented in the following section predict a possible contamination by HCl, a byproduct issued from precursor chemisorption. This led us to introduce a field of occupation HCl to account for its presence near a hafnium center, crystalline or not. By convention, for the hydrolysis phase, adsorbed water will be represented by the value 3 in the metal field of occupation. Each element of our table of atomic configuration will thus be a structure containing several fields of occupation (metal, Cl, OH, $O_a$, $O_b$, HCl) and being able to take various values. This occupancy table will play a major role during the execution of the software: in permanent evolution, it will be questioned and in turn modified very frequently until the final configuration is obtained. The use of grid indices, preset sites within a basic cell and fields of occupation will allow a faithful representation of the system, while guaranteeing a light and fast data-processing, authorizing an exploration on a mesoscopic scale.

*d. Neighborhood.* To determine, in an efficient way, if a chemical reaction is possible or not within the crystal, it is necessary to know the state of occupation of neighboring cells in the grid. Therefore, it is very important to know precisely the neighbors of a given site. We define 12 neighbors per lattice site, taking periodic boundary conditions into account.

*e. Substrate Modeling.* Figure 4 shows connection between the atomic model of network based configuration and a hydroxylated $Si/SiO_2$ (100) substrate, by the superposition of the chart of the system and the symbols of the network based model (Si−H are not represented). It is seen here that the silicon atoms take the places that are reserved for them (triangles) with internal displacements, in particular, on the level of dimers, being neglected. Siloxanes and hydroxyls are distributed according to options described previously. One notes that preset sites are empty, "on standby" for the arrival

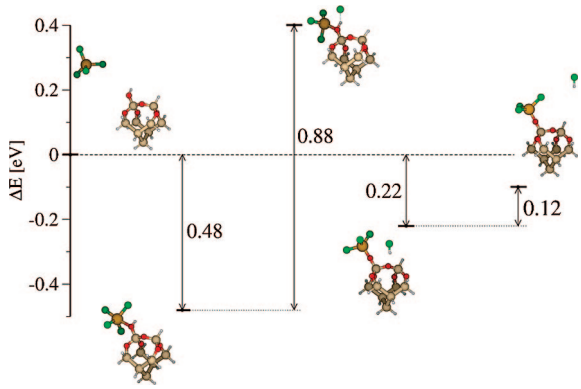



***Figure 5.*** Initial reaction path way and associated barriers in the case of HfCl4 based precursor ALD on $SiO_2$/Si.

of a hafnium or an oxygen. These crystalline positions will be occupied through molecular transient states, as considered in the fields of occupation of the table of configurations.

*2.2.1.3. Mechanisms and Events.* To avoid confusion, we need to define what we mean by "reaction step" and by KMC "event". A reaction step is an elementary or single chemical reaction. An event is related to the Monte Carlo procedure; it stands for a mechanism that occurs at a particular time on a well-determined site within the grid.

*a. Mechanism Definition.* The elementary reaction steps that constitutes the basis of any KMC approach can emanate from literature data, ab initio calculations, or a KMC investigation itself (in particular cases where the code is ready for validation procedures). As an example, Figure 5 illustrates precursor chemisorption, followed by HCl desorption. The calculated activation energies, using ab initio DFT, can serve as preliminary input values for the kinetic Monte Carlo mechanisms. The results of ab initio calculations are summarized in Section 3.1. Further, ALD growth is complicated by other reaction steps, not just the precursor molecule's adsorption and hydrolysis. An example is the densification (Table 1). Indeed, a key technological question that arises during the film growth is the necessary phase transition of the interesting materials from their molecular structure in the gas-phase precursors to their solid-state structure in the deposited film. This is particularly true for metallic compounds, such as oxides ($HfO_2$), where the metal has a covalent bonding structure in the precursor, with a small coordination number, while the metal oxides have mainly ionic structures with large coordination numbers. However, the complexity of the densification process is such that our present KMC implementation includes no changes in Hf or O coordination number. Further, because of the intrinsic difficulty in consideration of reaction steps related to these phenomena, the set up of a DFT modeling strategy becomes tricky. Later on in the paper, densification will be clarified, and we will show how this problem can be overcome in the KMC.

*b. Events Filtering and Modified Probability Sites.* Let us begin with an example, the precursor arrival mechanism. Obviously, there are lots of precursor arrival events, onto every surface site. But, at a given time, not all these events are possible. They must be filtered, or in other words, authorized or forbidden. The authorized events list is updated

***Table 1.*** List of Mechanisms Considered in the KMC Simulations, with Their Main Characteristics, As Described in the Text

| DFT investigation | KMC investigation[a] |
|---|---|
| 01 HfCl4 adsorption | 09 Dens. Inter_Cl_1N_cOH-iOH (all *k*) |
| 02 H2O adsorption | 10 Dens. Inter_Cl_1N_cOH-iCl (all *k*) |
| 03 HfCl4 Desorption | 11 Dens. Inter_Cl_1N_cCl-iOH (all *k*) |
| 04 HCl incorporation | 12 Dens. Inter_Cl_2N_cOH-iOH (all *k* except 2) |
| 05 H2O Desorption | 13 Dens. Inter_Cl_2N_cOH-iCl (all *k* except 2) |
| 06 H2O incorporation | 14 Dens. Inter_Cl_2N_cCl-iOH (all *k* except 2) |
| 07 back reaction | 15 Dens. Intra_Cl_1N_cOH-iOH ($k = 2$) |
| 08 HCl Desorption | 16 Dens. Intra_Cl_1N_cOH-iCl ($k = 2$) |
| | 17 Dens. Intra_Cl_1N_cCl-iOH ($k = 2$) |
| | 18 Dens. Intra_CC_1N_cOH-cOH ($k = 2$) |
| | 19 Dens. Intra_CC_1N_cOH-cCl ($k = 2$) |
| | 20 Dens. Intra_CC_2N_cOH-cOH ($k = 2$) |
| | 21 Dens. Intra_CC_2N_cOH-cCl ($k = 2$) |
| | 22 Dens. Bridge_Tl_2N_tOH-iOH ($k = 2$) |
| | 23 Dens. Bridge_Tl_2N_tOH-iCl ($k = 2$) |
| | 24 Dens. Bridge_Tl_2N_tCl-iOH ($k = 2$) |
| | 25 Dens. Bridge_Tl_3N_tOH-iOH ($k = 2$) |
| | 26 Dens. Bridge_Tl_3N_tOH-iCl ($k = 2$) |
| | 27 Dens. Bridge_Tl_3N_tCl-iOH ($k = 2$) |
| | 28 Dens. Bridge_TC_3N_tOH-cOH ($k = 2$) |
| | 29 Dens. Bridge_TC_3N_tOH-cCl ($k = 2$) |
| | 30 Dens. Bridge_TC_3N_tCl-cOH ($k = 2$) |
| | 31 Dens. Bridge_TC_4N_tOH-cOH |
| | 32 Dens. Bridge_TC_4N_tOH-cCl |
| | 33 Dens. Bridge_TC_4N_tCl-cOH |
| | 34 Dens. Bridge_TT_3N_tOH-tOH ($k = 2$) |
| | 35 Dens. Bridge_TT_3N_tOH-tCl ($k = 2$) |
| | 36 Dens. Bridge_TT_4N_tOH-tOH |
| | 37 Dens. Bridge_TT_4N_tOH-tCl |
| | 38 Dens. Bridge_TT_5N_tOH-tOH |
| | 39 Dens. Bridge_TT_5N_tOH-tCl |

[a] Densifications may result from reactions between Cl or OH terminals in ionic (i), cluster (c), or tree (t) configurations.

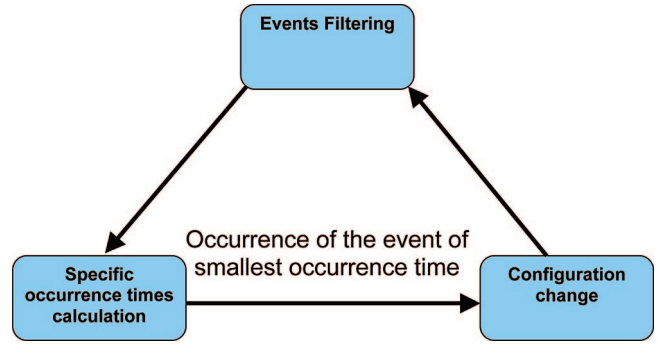



***Figure 6.*** Kinetic Monte Carlo cycle.

at each system configuration change. Moreover, if one given event is forbidden and becomes authorized, its "specific time" is also updated (see Section 2.2.1.2b for details about this point). In contrast, if the event is authorized and remains authorized, its "specific time" is kept as is. In any other case, the forbidden event's "specific time" is set to a value greater than the overall experiment duration. If not reauthorized again, this event will never occur. To save computing time, the filtering procedure is only called when necessary by a "smart call" routine, Figure 6. Another way to save computing time is to consider modified probability sites (MPS). In this procedure, after the occurrence of a given event, one considers only a local reactualization of the sites/events/time of occurrence. Rigorously, all the sites/events could be reactualized, which would require a huge amount of computing time. This procedure is thus used on the basis of the

concept that an event has a limited impact in space and that domains far from the occurring event do not see it.

## 3. Results and Discussion

**3.1. Ab Initio Results.** To date, almost all ab initio investigations have focused on the chemistry of possible reactions, during both precursor and water exposures of the ALD of $HfO_2$. These predictive methods, cluster or periodic based-DFT, are particularly suitable for chemical mechanisms that are difficult to reach by experiment or other modeling strategies.

Here, the reactions between the gaseous precursors $HfCl_4$ or $H_2O$ with the hydroxylated $SiO_2$ surface are predicted to proceed via an exchange mechanism[65,66] and can be separated into two half reactions. The first half-cycle reaction (1) takes place during the initial step of atomic layer deposition of $HfO_2$, that is, the decomposition of $HfCl_4$ precursor molecules, followed by a purge

$$||SiO_2-OH + HfCl_4 \rightarrow ||SiO_2-O-HfCl_3 + HCl \quad (1)$$

The second half cycle (2) is exposure to water molecules and is also followed, in practice, by a purge period.

$$||SiO_2-O-Hf(OH)_{3-x}-Cl_x + H_2O \rightarrow SiO_2-O-Hf(OH)_{4-x}-Cl_{x-1} + HCl \quad (2)$$

where $x$ has values of 1−3.

First, we give the results of the first half-cycle reaction at the initial deposition step. The second task is dedicated to the second half reaction. We present results and compare them with those related to $ZrO_2$. Data from the literature are also discussed. However, the basic mechanisms that are given in the following are crucial to understand the results all along the multiscale procedure up to the KMC/experiment relation. All basic mechanisms have been introduced in the KMC and are activated all along the ALD process duration, except chemisorptions of water and precursor from gas-phase exposure that are limited to specific semicycles. Their desorptions, reactions, and backreactions are possible through all cycles. Their occurrence will depend on the computed activation energies.

*3.1.1. Precursor/Surface Reaction: Overview of Existing Results.* Geometrical structures and energetic diagram of the total system along the decomposition reaction pathways of $HfCl_4$ on the $SiO_2$ substrate are presented in Figure 5. A chemisorbed state is found where Hf/O interaction is occurring without substantial modification of the precursor integrity (0.48 eV adsorption energy). This adsorption energy is weaker than those obtained by Musgrave and co-workers.[66] This difference is the result of the differences in cluster geometry and basis set. Both parameters are more extended in our study. A dissociation path is then calculated, and we obtain an activation barrier of 0.88 eV. This barrier is higher than the 0.7 eV barrier obtained by Widjaja et al..[68] The difference is caused by the methodology used in the search of transition state and the surface models that are different. Beyond the saddle point, HCl molecule formation is observed during minimization with subsequent adsorption of this molecule on

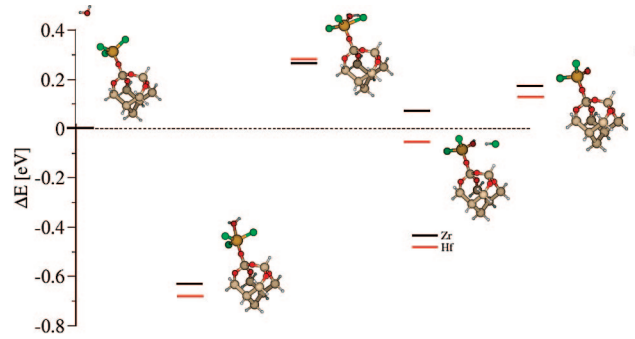



***Figure 7.*** Reaction path and predicted energetics for the reactions of $H_2O$ and $SiO_2-O-HfCl_3$ and $H_2O$ and $SiO_2-O-ZrCl_3$.

top of the dimer inserted oxygen atom. Globally, it is expected from these results that the deposition with the desired coverage of a complete monolayer of precursors is a difficult task. The sticking of the precursor is low. The high activation barrier for decomposition is compared to the precursor non dissociative chemisorption. Moreover, the dissociative incorporation is reversible. The dissociated precursor molecule is itself thermodynamically unstable with respect to the chemisorbed state. This authorizes the precursor to recombine and to be rejected from the surface. Fortunately, the HCl byproduct desorption is fast, because of its low activation energy of only 0.12 eV, and inhibits further reverse reaction. These qualitative aspects will be confirmed at the KMC modeling level.

*3.1.2. Hydrolysis Reaction.* The second half cycle, which introduces $H_2O$ into the reaction chamber, is intended to regenerate the surface-OH group by reaction 2. The reaction path and predicted energetic for the reaction of $H_2O$ on $||SiO_2-O-HfCl_3$ sites are presented in Figure 7. For comparison, we also report, in this figure, the results of the reaction of $H_2O$ on $||SiO_2-O-HfCl_3$. The energy levels are displayed relative to initial state energy. Further, Table 2 summarizes the energetic data for the two reactions and compares them with corresponding results obtained by Musgrave and co-workers.[67−69]

The quantitative energetic of the reaction 2, shown in Figure 8, shows some similarities between $HfO_2$ and $ZrO_2$. Indeed, water molecule adsorbs onto the $HfCl_3$ (or $ZrCl_3$) covalently bonded to the surface by donating a lone pair from the oxygen atom to the empty d shell of the Hf. In the case of HF (or Zr), the complex chemisorbed state is 0.68 eV (or 0.62 eV) more stable than the initial state. After the adsorption of $H_2O$, the reaction proceeds via the transition state shown in Figure 8. In the transition state, O in $H_2O$ is bonded to Hf with a bond length of 1.9 Å. The Hf−Cl distance is increased to 3.8 Å, which indicates that the bond is broken in the transition state. We also note that the interaction between the O lone pair and the HCl antibonding orbital leads to the delocalization of the lone pair electrons, which indicates the partial breaking of the O−H bond and the partial formation of HCl bond. The reaction is exothermic by 0.19 eV with an activation barrier of 0.97 eV. As shown in Table 2 and Figure 8, the second half reaction for Zr has the same behavior as for Hf. Indeed, the saddle point energies for the two systems are also close leading to an activation

***Table 2.*** Comparisons of Energetic (In eV) for the Si−O−M(Cl)$_3$+ $H_2O$ (M = Hf, Zr) and Those Obtained from Literature[a]

| | M−$Cl_3$−$H_2O$ complex | TS[b] | M−$OHCl_2$−HCl complex | HCl desorption |
|---|---|---|---|---|
| M = HF | 0.68 | 0.97 | 0.63 | 0.20 |
| M = Zr | 0.62 | 0.90 | 0.69 | 0.09 |
| M = Zr (results of Musgrave, ref 68) | 0.78 | 0.88 | 0.75 | 0.05 |

[a] The energies of M−$Cl_3$−$H_2O$ are relative to the entrance channel. The energies of TS, M−$OHCl_2$−HCl complex are relative to the M−$Cl_3$−$H_2O$ complex. The energies of HCl desorption are relative to the M−$OHCl_2$−HCl complex. [b] Transition state.

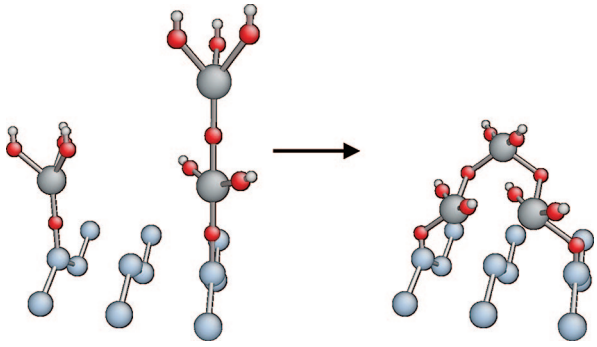


***Figure 8.*** Example of densification mechanism: multilayer noncrystalline/tree.

barrier of 0.97 and 0.90 eV for Hf and Zr, respectively. In comparison with the data obtained by Musgrave et al., the choice of cluster geometry and basis set led to systematic differences in relative energies, except that a similar barrier is obtained by Musgrave (0.88 eV) and us (0.90 eV).

Further, on the basis of the results obtained by Musgrave and us, the use of clusters models should be sufficient to determine the reaction mechanisms. To determine the credibility of DFT in predicting these mechanisms, Heyman et al.[35] reported the calculations (especially transition states) on different precursors using DFT, MP2, and QCISD(T) and concluded that both ab initio approaches resulted in the same transition states that were found using DFT (B3LYP).

*3.1.3. Densification Mechanisms.* Beyond reaction mechanisms appears a key issue for a better understanding of high-*k* film growth, we will call the phenomenon "densification". In the gas phase, metallic precursors exhibit a covalent molecular structure characterized by a small coordination number, whereas the resulting deposited metallic oxide has an ionic type of structure characterized by a large coordination number. This necessary phase transition has been poorly investigated to date in literature.[50] In our recent work, we show that this chemical process is mediated by the presence of oxygen atoms: sharing them locally allows a local redistribution of oxygen atoms and metallic centers to operate the desired transition. This implies an increase of the coordination of both Hf and oxygen atoms and a "more dense" local reallocation of the atoms through the collapse of dendritic tree like structures into densely packed oxide films. The dendritic structures result from an accumulation of adsorbed molecules, preventing a dense structure by steric hindrance. However, the complexity of the densification process is such that our KMC model still poorly describe this phenomenon leading to no increase of the coordination number.

In addition, we expect new insight from DFT calculations in this area in the future (currently under investigation in our team) because no such reaction pathway has been calculated in the frame of the ALD process specifically.

**3.2. Preliminary KMC Results.** *3.2.1. Mechanisms in Kinetic Monte Carlo Code.* The kinetic Monte Carlo code contains all the above generic mechanisms, declined in various cases and configurations. These are detailed in Table 1, where we distinguish between two types of mechanisms, respectively listed in the two columns of Table 1. The first column contains eight mechanisms whose thermodynamic and kinetic parameters are calculated using DFT. It contains precursor adsorption, desorption, incorporation and back reaction, $H_2O$ adsorption, desorption and incorporation (hydrolysis), and HCl byproduct desorption.

The second column contains 31 (09−39) mechanisms of densification type whose rates are empirically adjusted by comparison of simulation results with experimental data. These densification mechanisms can occur between precursors situated on the same layer (intra) or different layers (inter). They can also occur in situations where the creation of a bridge, with overhangs left in the growing film, is necessary. On the other hand, densifications can occur between neighboring sites, up to the fifth neighbors. Densifications may also result from reactions between Cl or OH terminals in ionic (i), cluster (c), or tree (t) configurations. These denominations refer respectively to terminals directly adsorbed on the film surface (i), which are therefore in ionic positions, to terminals on precursor molecule deposited on the film surface as a cluster (c), and to terminals on treelike precursor molecules (t). These latter are molecules bonded to an already adsorbed molecule. They allow the adsorption of several layers of precursor molecules on the surface.

All the above information, together with the layer number *k*, where the corresponding events can occur, is reported in Table 1. In addition, in Figure 8, we give an example of densification mechanism.

The energetic of reactions 09−39 are set to one arbitrary value and later fitted to experiment. This will help to bring out the importance of further modeling of these reactions.

*3.2.2. Optimization of Substrate Hydroxylation.* Hydroxylated surfaces can be prepared using various processes[70,71] that allow preparation of surfaces with different densities of reaction sites. We thus simulate a first ALD phase, at 300 °C, with a duration of 100 ms verusus the 50 ms duration generally used in experiments. We have performed simulations ((10 × 10) atoms substrates) on five different initial substrates containing various OH concentrations. The first four cases concern 25%, 50%, 75%, and 100% hydroxyl concentrations, with siloxane bridges distributed randomly; 100% coverage corresponds to one OH per surface silicon atom. The fifth case concerns the simulation of the hydroxy-

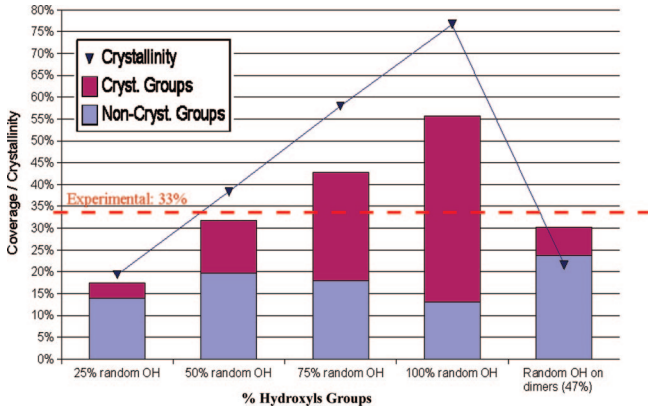



***Figure 9.*** Evolution of coverage rates ($HfO_2$ coverage) and crystallinity with different initial substrates.

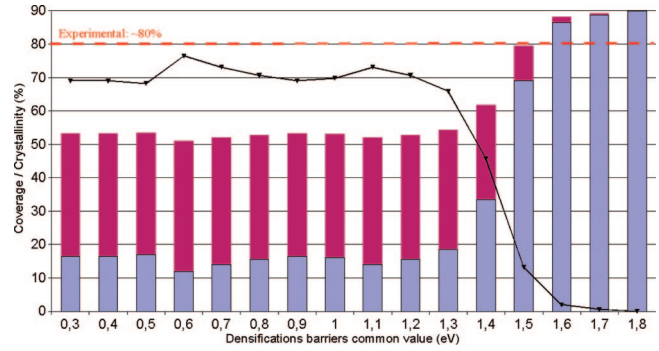



***Figure 10.*** Calibration of energies of the first densifications.

lation process with the option "no more than one OH on each dimer". This process leads to 47% hydroxyl coverage.

The coverage rates ($HfO_2$ coverage) obtained, after one ALD cycle, are represented on the graph of Figure 9, where we have also distinguished between the crystalline and molecular states of Hf. Again, 100% coverage corresponds to one Hf per surface silicon atom. The triangles in Figure 9 show the crystallinity ratio. An increase in the total coverage is observed when one increases the number of reactive sites on the substrate. Moreover, crystallinity follows the same tendency: this is the result of the possibility for a chemically absorbed precursor to bridge itself if an OH is available in its vicinity. In the case with 100% hydroxyl coverage of the surface, it is noted that the total coverage amounts to approximately 55%, indicating that a majority of hafnium atoms are thus bridged. In this case, each Hf reacts with two OH: one for the sticking and one for the bridging process. The bridging process can be considered as one type of densification mechanism because it allows the elimination of one branch within the molecular tree structure.

We cannot judge yet if this is realistic because the barriers of these first densifications are unknown, and arbitrary values are introduced at this stage. At the end of the first ALD cycle performed at 300 °C on a chemical silicon oxide, a total coverage of 35% has been experimentally measured.[69] This value is in good agreement with our simulations with 50% OH and 47% OH. These concentrations are indeed the most probable ones when referring to the model of Zhuravlev, dealing with the hydroxylation of $SiO_2$ surfaces.[72] In addition, in the last case, by limiting the local OH concentrations, we hinder the densification mechanisms: crystallinity is 22% in the fifth case, while 38% in the second case with 50% OH.

*3.2.3. Calibration of Energies of First Densifications.* In practice, densification mechanisms in the kinetic Monte Carlo account for all transitions from molecular state to ionic or crystalline states, that is, the nodes of the lattice. These include reactions between grafted molecular ligands and the surface, such as the bridging process (eq 3) or reactions between two grafted ligands close to each other. The activation energies of all the mechanisms of densification are unknown and not well understood to date.[73] We have arbitrarily fixed them at 0.5 eV for preliminary simulations. Their values have no incidences on the coverage at low hydroxyl concentrations and at the usual growth temperatures. But, if underestimated, the coverage can decrease at high hydroxyl concentrations. In this case, the most probable reaction is between two first neighbors OH, to bridge a single Hf, as expressed in eq (3)

$$\mathrm{Si{-}O{-}HfCl_3 + Si{-}OH \rightarrow Si{-}O{-}HfCl_2{-}O{-}Si + HCl} \quad (3)$$

Actually, in the KMC scheme, the global bridging mechanism is broken in two steps, chemisorption and densification, each one consuming a single OH. If the energy barriers for the first densification are increased, the densification rates decrease, and the most probable configurations will be two separate chemisorptions on two OH groups, without any densification, as in eq 4

$$\mathrm{2Si{-}OH + 2HfCl_4 \rightarrow 2Si{-}O{-}HfCl_3 + 2HCl} \quad (4)$$

The competition between eqs 3 and 4 can be quantified by comparing KMC results with the experimental values of the coverage.[70] From these experiments, we can estimate very roughly that with a surface covered with 90% hydroxyl, one should obtain, after one ALD cycle on a (10 × 10) atom substrates, a $HfO_2$ coverage of approximately 80%. This is in total disagreement with preceding simulations. We thus launched several simulations of the first precursor injection phase of ALD, performed at 300 °C during 200 ms, while varying the energy barrier for densifications between 0.3 and 1.8 eV. We should mention that in a given simulation, the same value is applied to all the densifications. Figure 10 shows the $HfO_2$ coverage and crystallinities obtained for these various densification barriers. For weak densification barriers, we observe a strong crystallinity, proof that a majority of Hf atoms are bridged thanks to the great number of hydroxyls. For higher values of densification barriers, chemisorption takes place, leading then to less crystallinity but more extensive coverage. The simulation that approaches 80% of experimental measured coverage is that with a densification barrier of 1.5 eV. This calibration should be, at first sight, very approximate. The experimental criterion translated into atomic terms is not inevitably reliable, and all the densifications may not be equivalent. Despite these naive considerations, the transition seen by variation of the densification energy appears to be abrupt, a few tenths of

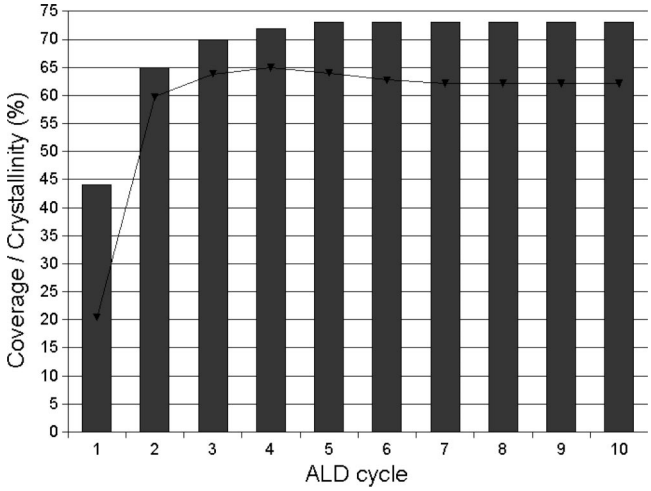



***Figure 11.*** Evolution of coverage rates ($HfO_2$ coverage) and crystallinity with respect to ALD cycles number. The coverage is cumulative. The legend is the same as in Figure 10.

an electronvolt. This suggests that the calibration of a mechanism, or a class of mechanisms, is accessible through KMC by confrontation to experiment.

Waiting for subsequent quantum contributions or experiment investigations, we adopt this single value of 1.5 eV as a first approximation for all the densification mechanisms. We hope that it may help and shall be used as a basis for DFT investigations, which can now allow us to distinguish between different types of densification mechanisms (listed in Table 1), associated with different activation barriers, and to determine the most efficient ones. This proves that the simulator can lend itself to a calibration on experimental data, even in a very approximate way but still necessary to get insights toward complex physicochemical issues.

*3.2.4. Kinetics of Growth.* Until now, we have validated the basics of the method and the first mechanisms introduced in the KMC scheme, without exceeding the first phase of the first ALD cycle. We now will try to go beyond this stage by carrying out simulations on several ALD cycles under standard conditions: at 300 °C and 1.33 mbar pressure for the two precursors. The phases of injection will last 50 ms, as in experiments, and are followed by a 4 s of purge.

In experiments, several modes of growth are successively observed.[70,71] After the first cycle, consuming a great number of metal precursors, a slow transient state is established to complete the coverage of the substrate. Then, a steady operation starts, faster, where $HfO_2$ is deposited on $HfO_2$ substrate. By using atomic scale KMC, we try to reproduce all these aspects of the kinetics of growth on ($10 \times 10$) and ($20 \times 20$) atoms substrates during tens of ALD cycles.

We first launched a simulation of 10 ALD cycles, whereas the mechanism of opening of a siloxane bridge was not yet implemented in the software package. Indeed, it was experimentally observed that that the opening of siloxane bridges only took place at higher temperatures.

Figure 11 shows the evolution of the cumulated coverage ($HfO_2$ coverage) and the total crystallinity. We note a great number of incorporations during the first cycle leading to 44% coverage. This value is slightly larger than experimental observations (35%), undoubtedly because of an overevaluation of the hydroxyl concentrations given by the Zhuravlev model[72] Then, from cycle 2 to 5, the coverage saturates gradually until a maximum of 73% is reached. Beyond cycle 5, the growth is stopped. This is caused by the inability of the densification mechanisms to be fully efficient. An examination of the final structure indicates specific sites where further densification mechanisms should occur but cannot be performed by the mechanisms already implemented in our software package. We are currently considering rewriting the algorithm used to treat densification mechanisms to overcome this problem. The rate of crystallinity follows the same tendency and is stabilized around 62%, meaning that the densification had taken place, but not fully, because of, at least, an unsatisfactory description of densification. However, one cannot be satisfied with this situation because the experiments show that the growth continues, certainly slowly, after the strong coverage observed after the first cycle.[70,71] An opening of the siloxane bridges by a water molecule has been suggested that could allow the total coverage to be reached. The opening of the siloxane bridges is postponed during the hydrolysis phases and could bring additional reactive sites on the initial surface. However, no consensus exists to explain how the $HfO_2$ coverage is reached exactly. Currently, effort is undertaken on this subject through the improvement of the densification mechanisms suggested recently by DFT investigations.[73]

## 4. Conclusions

This work presents an original kinetic Monte Carlo algorithm, developed within a multiscale strategy: from molecule−surface interactions treated mostly at the DFT level of modeling, to atomic scale film growth performed via a KMC technique. We have given details on how ab initio calculations can help in the identification of the relevant elementary reaction mechanisms, before their definite implementation into the Monte Carlo simulator that can handle millions of atoms during seconds of simulation runs. We then show how to develop a KMC method aimed at dealing with a promising experimental deposition procedure: the ALD. This technique is expected to result in a unique change in microelectronics history by replacing traditional $Si/SiO_2$ interface by a deposited high-*k* material onto silicon. The basics of a specific KMC have been reported, and preliminary examples for the validation of the overall multiscale strategy are detailed. We have shown how KMC provides a unique and fundamental understanding of the growth mechanism and growth kinetics of the dielectric ($HfO_2$) onto silicon. The relation of KMC with experiment is lighted. We have also shown that KMC, allowing a global view of mesoscopic structures, may shed light on shortcomings in the model. New mechanisms can then be considered for ab initio DFT investigations. The relevant example here has been the case of densification mechanisms. We hope that this work will bring new advances for better control of the $HfO_2/SiO_2/Si$ systems and for further optimization of the processing parameters with the aim of integrating hafnium oxide in the heart of future generations of MOS device. Beyond this targeted application, we believe that the development of the

KMC, detailed here, is generic and may serve other new concepts in nano- and bionanotechnologies.

**Acknowledgment.** C.M. acknowledges the "Conseil regional de la Martinique" for the funding.

## Appendix

**A**. **Temporal Dynamics.** Kinetic Monte Carlo approaches differ from static Monte Carlo techniques by the introduction of time. Most of the Monte Carlo procedures used in condensed matter physics are static (the Metropolis algorithm for instance), and their interest concerns the equilibrium structure properties. KMC is totally different in philosophy. Here, the Monte Carlo technique is introduced for the stochastic aspects, in connection with the time incremental procedure, that is, it is not always the most probable event that occurs. Different algorithms do exist for KMC. We suggest in the following a procedure that becomes efficient, with regard of the more conventional BKL algorithm, when dealing with many different events related to the same mechanism, but having different probabilities. This means that the probability depends also on the site of occurrence. This is the case of systems where local deformation energy is taken into account (each event has therefore a different probability of occurrence) and where the system exhibits a complex chemistry.

**A1**. **Atomic Layer Deposition (ALD).** The KMC software discussed here is built to simulate the ALD experimental process. ALD consists of four phases: precursor pulse, precursor purge, hydrolysis, and water purge. Each one has its own thermodynamic parameters and duration.. Some mechanisms are always possible, whereas others can only occur in a typical phase. For instance, "precursor arrival" will obviously exist during the first phase only. Furthermore, the complete ALD process consists of several cycles of these four phases.

**A2**. **Time Management.** When the software is launched, after preliminary initializations (parameters, mechanisms, neighborhood, configuration), it makes a global scan to determine which events are authorized and which are forbidden. Authorized ones get a "specific occurrence time": this is the time that the considered event typically takes to occur as soon as it becomes authorized. Forbidden ones get the maximum time so that they can never occur.

The "specific occurrence time" of the event $m$ on site ($i$, $j$, $k$) is given by

$$T_{i,j,k,m} = \frac{-\log(Z)}{\lambda_m}$$

where $Z$ is a random number uniformly distributed between 0 and 1 and $\lambda_m$ is the probability, per unit time, of occurrence of the mechanism. It is expressed as

$$\lambda_m = \nu \exp\left(\frac{\Delta E_m}{k_B T}\right)$$

where $\nu$ is on the order of the typical lattice vibration frequency, $\Delta E_m$is the activation energy of the mechanism, $k_B$ the Boltzmann constant, and $T$ the temperature. The two arrival mechanisms (1 precursor arrival and 2 water arrival) obey Maxwell−Boltzmann statistics of the gas phase and have different probability expressions

$$\lambda_{1,2} = \frac{\mathrm{Cst}PS}{\sqrt{M_{1,2}T}}$$

where Cst is a constant, $P$ the pressure, $S$ the elementary 2D-cell area, $M_{1,2}$ the molar mass of the considered species, and $T$ the temperature. In this scheme, we have a huge list of "specific times", a kind of "calendar", just as if all events that will occur were already foreseen. Not exactly in fact, because this calendar will often be updated as the configuration evolves. As soon as the calendar is updated, the software finds the minimum "specific occurrence time" and the corresponding event occurs. After this minimum time has passed, it is withdrawn from the other times contained in the "calendar". Then, the configuration is edited and the pertinent events (that are likely to become authorized) are filtered. We then go back to the first KMC stage where the minimum time is searched.

## References

(1) Packan, P. A. *Science.* **1999**, *285*, 2079.

(2) Wilk, G. D.; Wallace, R. M.; Anthony, J. M. *J. Appl. Phys.* **2001**, *89*, 5243.

(3) Wilk, G. D.; Muller, D. A. *Appl. Phys. Lett* **2003**, *83*, 3984.

(4) (a) George, S. M.; OTT, A. W.; Klaus, J. W. *J. Phys. Chem.* **1996**, *100*, 13121. (b) Xu, Z.; Houssa, M.; Gendt, S. D.; Heyns, M. *Appl. Phys. Lett.* **2002**, *80*, 1975.

(5) Ferrari, S.; Scarel, G.; Wiemer, C.; Fanciulli, M. *J. Appl. Phys.* **2002**, *92*, 7675.

(6) Park, H. B.; Cho, M. J.; Park, J.; Lee, S. W.; Hwang, C. S.; Kim, J. P.; Lee, J. H.; Lee, N. I.; Kang, H. K.; Lee, J. C.; Oh, S. J. *J. Appl. Phys.* **2003**, *94*, 3641.

(7) Kawahara, T.; Torii, K. *IEICE Trans. Elect.* **2004**, *E87C* (1), 2–8.

(8) Ganem, J. J.; Trimaille, I.; Vickridge, I. C.; Blin, D.; Martin, F. *Nucl. Instrum. Methods Phys. Res., Sect. B* **2004**, *219*, 856.

(9) Green, M. L.; Ho, M. Y.; Busch, B.; Wilk, G. D.; Sorsch, T.; Conard, T.; Brijs, B.; Vandervorst, W.; Raisanen, P. I.; Muller, D.; Bude, M.; Grazul, J. *J. Appl. Phys.* **2002**, *92*, 7168.

(10) Triyoso, D. H.; Hegde, R. I.; Grant, J.; Fejes, P.; Liu, R.; Roan, D.; Ramon, M.; Werho, D.; Rai, R.; La, L. B.; Baker, J.; Garza, C.; Guenther, T.; White, B. E.; Tobin, P. J. *J Vac. Sci. Technol. B.* **2003**, *22*, 2121.

(11) Rittersma, Z. M.; Roozeboom, F.; Verheijen, M. A.; van Berkum, J. G. M.; Dao, T.; Snijders, J. H. M.; Vainonen-Ahlgren, E.; Tois, E.; Tuominen, M.; Haukka, S. *J. Electrochem. Soc.* **2004**, *151*, 716.

(12) Chang, H. S.; Hwang, H.; Cho, M. H.; Moon, D. W. *Appl. Phys. Let.* **2005**, *86*, 031906.

(13) Triyoso, D. H.; Ramon, M.; Hegde, R. I.; Roan, D.; Garcia, R.; Baker, J.; Wang, X. D.; Fejes, P.; White, B. E.; Tobin, P. J. *J. Electrochem. Soc.* **2005**, *152*, 203.

(14) Cho, M.; Park, H. B.; Park, J.; Lee, S. W.; Hwang, C. S.; Jeong, J.; Kang, H. S.; Kim, Y. W. *J. Electrochem. Soc.* **2005**, *152*, F49.

(15) Kukli, K.; Aaltonen, T.; Aarik, J.; Lu, J.; Ritala, M.; Ferrari,

S.; Harsta, A.; Leskela, M. *J. Electrochem. Soc.* **2005**, *152*, F75.

(16) Triyoso, D. H.; Hegde, R. I.; White, B. E.; Tobin, P. J. *J. Appl. Phys.* **2005**, *97*, 124107.

(17) Hellin, D.; Delabie, A.; Puurunen, R. L.; Beaven, P.; Conard, T.; Brijs, B.; De Gendt, S.; Vinckier, C. *Anal. Sci.* **2005**, *21*, 845.

(18) Kirsch, P. D.; Quevedo-Lopez; MA; Li, H. J.; Senzaki, Y.; Peterson, J. J.; Song, S. C.; Krishnan, S. A.; Moumen, N.; Barnett, J.; Bersuker, G.; Hung, P. Y.; Lee, B. H.; Lafford, T.; Wang, Q.; Gay, D.; Ekerdt, J. G. *J. Appl. Phys.* **2006**, *99*, 023508.

(19) Park, I. S.; Lee, T.; Choi, D. K.; Ahn, J. *J. Korean Phys. Soc.* **2006**, *49*, S544.

(20) Cho, M. J.; Degraeve, R.; Pourtois, G.; Delabie, A.; Ragnarsson, L. A.; Kauerauf, T.; Groeseneken, G.; De Gendt, S.; Heyns, M.; Hwang, C. S. *IEEE Trans. Electron Devices* **2007**, *54*, 752.

(21) Nyns, L.; Hall, L.; Conard, T.; Delabie, A.; Deweerd, W.; Heyns, M.; Van Elshocht, S.; Van Hoornick, N.; Vinckier, C.; De Gendt, S. *J. Electrochem. Soc.* **2006**, *153*, F205.

(22) Ritala, M.; Leskela, M.; Niinisto, L.; Prohaska, T.; Friedbacher, G.; Grasserbauer, M. *Thin Solid Films* **1994**, *250*, 72.

(23) Aarik, J.; Aidla, A.; Kiisler, A. A.; Uustare, T.; Sammelselg, V. *Thin. Solid Films* **1999**, *340*, 110.

(24) Kukli, K.; Ihanus, J.; Ritala, M.; Leskela, M. *Appl. Phys. Lett.* **1996**, *68*, 3737.

(25) Aarik, J.; Aidla, A.; Kiisler, A. A.; et al. *Thin Solid Films* **1999**, *340*, 110.

(26) Aarik, J.; Aidla, A.; Mandar, H. *J. Cryst. Growth* **2000**, *220*, 105.

(27) Aarik, J.; Aidla, A.; Kikas, A. *Appl. Surf. Sci.* **2004**, *230*, 292.

(28) Kukli, K.; Aarik, J.; Uustare, T. *Thin Solid Films* **2005**, *479*, 1.

(29) de Almeida, R. M. C.; Baumvol, I. J. R. *Surf. Sci. Rep.* **2003**, *49*, 1.

(30) Wallace, R. M.; Wilk, G. D. *Crit. Rev. Solid State Mater. Sci.* **2003**, *28*, 231.

(31) Wang, D.; Wang, Q.; Javey, A. *Appl. Phys. Lett.* **2003**, *83*, 2432.

(32) Wong, H.; Iwai, H. *Microelectron. Eng.* **2006**, *83*, 1867.

(33) Aarik, J.; Aidla, A.; Sammelselg, V.; Siimon, H.; Uustare, T. *J. Cryst. Growth* **1996**, *169*, 496.

(34) Puurunen, R. L. *Chem. Vap. Deposition.* **2005**, *11*, 79.

(35) Heyman, A.; Musgrave, C. B. *J. Phys. Chem. B.* **2004**, *108*, 5718.

(36) Samantaray, C. B.; Sim, H.; Hwang, H. *Appl. Surf. Sci.* **2005**, *242*, 121.

(37) Elliott, S. D.; Pinto, H. P. *J. Electroceram.* **2004**, *13*, 117.

(38) Samantaray, C. B.; Sim, H.; Hwang, H. *Appl. Surf. Sci.* **2004**, *239*, 101.

(39) Mukhopadhyay, A. B.; Musgrave, C. B. *App. Phys. Lett.* **2007**, *90*, 173120.

(40) Forst, C. J.; Blochl, P. E.; Schwarz, K. *Comput. Mater. Sci.* **2003**, *27*, 70.

(41) Kawamoto, A.; Cho, K. J.; Dutton, R. *J. Comput.-Aided Mater. Des.* **2001**, *8*, 39.

(42) Musgrave, C. B.; Han, J. H.; Gordon, R. G. *Abstr. Pap. Am. Chem. Soc.* **2003**, *226*, U386.

(43) Mukhopadhyay, A. B.; Sanz, J. F.; Musgrave, C. B.; Han, J. H.; Gordon, R. G. *J. Phys. Chem. C.* **2007**, *26*, 9203.

(44) Ganem, J. J.; Trimaille, I.; Vickridge, I. C.; Blin, D.; Martin, F. *Nucl. Instrum. Methods Phys. Res., Sect. B* **2004**, *219*, 856.

(45) Mui, C.; Musgrave, C. B. *J. Phys Chem. B.* **2004**, *108*, 15150.

(46) Xu, Y.; Musgrave, C. B. *Appl. Phys. Lett.* **2005**, *86*, 192110.

(47) Mukhopadhyay, A. B.; Musgrave, C. B. *Chem. Phys. Lett.* **2006**, *421*, 215.

(48) Ren, J.; Zhang, Y. T.; Zhang, D. W. *THEOCHEM* **2007**, *803*, 23.

(49) Jeloaica, L.; Esteve, A.; Djafari Rouhani, M.; Esteve, D. *Appl. Phys. Lett.* **2003**, *83*, 542.

(50) Esteve, A.; Jeloiaca, L.; Mazaleyrat, G.; Dkhissi, A.; Djafari Rouhani, M.; Ali Messaoud, S.; Fazouan, N. *MRS Bull.* **2003**, *786*, 35.

(51) Jeloaica, L.; Esteve, A.; Dkhissi, A.; Djafari Rouhani, M. *Comput. Mater. Sci.* **2005**, *33*, 2005.

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

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

(54) Bergner, A.; Dolg, M.; Kuechle, W.; Stoll, H.; Preuss, H. *Mol. Phys.* **1993**, *80*, 1431.

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

(56) Schaefer, A.; Huber, C. F.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(57) Weldon, M.; Stefanov, B. B.; Raghavachari, K.; Chabal, Y. I. *Phys. Rev. Lett.* **1997**, *79*, 2851.

(58) Helgaker, T. *Chem. Phys. Lett.* **1991**, *182*, 503.

(59) (a) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165. (b) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571. (c) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V. B., A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A.; *Gaussian 03*, revision D.02; Gaussian,Inc.: Wallingford CT, 2004.

(61) Kotrla, M. *Comput. Phys. Commun.* **1996**, *97*, 82.

(62) Smilauer, P.; Vvedensky, D. D. *Phys. Rev. B.* **1995**, *52*, 14263.

(63) Schofield, W. C. E.; McGettrick, J. D.; Badyal, J. P. S. *J. Phys. Chem.B.* **2006**, *110*, 17161.

(64) Wyckoff, R., Ed. *Crystal Structure*; John Wiley & Sons: New York, 1965; Vol. 1, p 231.

(65) Ritala, M.; Leskela, N. in *Handbook of Thin Film Materials*; Nalwa, H. S., Ed.; Academic Press: New York, 2001; Vol. 1, p 57.

(66) Rahtu, A.; Ritala, M. *J. Mater. Chem.* **2002**, *12*, 1484.

(67) Widjaja, Y.; Musgrave, C. B. *J. Chem. Phys.* **2002**, *117*, 1931.

(68) Widjaja, Y.; Musgrave, C. B. *Appl. Phys. Lett.* **2002**, *81*, 304.

(69) Han, J. H.; Gao, G.; Widjaja, Y.; Garfunkel, E.; Musgrave, C. B. *Surf. Sci.* **2004**, *550*, 199.

(70) Blin, D. PhD thesis, University of Montpellier, 2003.

(71) Renault, O.; Samour, D.; Rouchon, D.; Holliger, Ph.; Papon, A. M.; Blin, D.; Marthon, S. *Thin Solid Films* **2003**, *428*, 190.

(72) Zuravlov, L. T. *Colloids Surf. A* **2000**, *173*, 1.

(73) Olivier, S.; Ducere, J. M.; Mastail, C.; Esteve, A.; Landa, G.; Djafari Rouhani, M. *Chem. Mater.* **2008**, *20*, 1555.

CT8001249

JCTC Journal of Chemical Theory and Computation

# Integral Equation Study of the Hydrophobic Interaction between Graphene Plates

Jesse J. Howard,[†] John S. Perkyns,[†] Niharendu Choudhury,[‡] and B. Montgomery Pettitt*,[†]

*Department of Chemistry, University of Houston, Houston, Texas 77204-5003, and Theoretical Chemistry Section, Bhabha Atomic Research Centre, Mumbai 400 085, India*

Received July 17, 2008

**Abstract:** The hydrophobic association of two parallel graphene sheets is studied using the 3D-RISM HNC integral equations with several theoretical methods for the solvent distribution functions. The potential of mean force is calculated to study the effects of the aqueous solvent models and methods on the plates as a function of distance. The results of several integral equations (IE) are compared to MD simulations for the same model. The 3D-IEs are able to qualitatively reproduce the nature of the solvent effects on the potential of mean force but not quantitatively. The local minima in the potential of mean force occur at distances allowing well defined layers of solvent between the plates but are not coincident with those found in simulation of the same potential regardless of the theoretical methods tested here. The dewetting or drying transition between the plates is generally incorrectly dependent on steric effects with these methods even for very hydrophobic systems without solute−solvent attractions, in contradiction with simulation.

## I. Introduction

Since the conception and implementation of the three-dimensional - reference interaction site model (3D-RISM) integral equation (IE) method, the theory has been applied to large anisotropic models to obtain structural details not possible using the analogous one-dimensional theory.[1−3] The solutions to the IEs are used to predict the thermodynamic properties of the solution as well as the spatial distribution of the solution species.[4,5] Studies of the conformational stability of molecular solutes based on thermodynamic considerations have shown the structural stability to depend on a delicate balance between competing forces among the solution species.[6] This is especially true for biomolecules where the folding mechanism and stability are determined by thermodynamic interactions within and between the solute and solvent which help to direct the folding of the native structure along a converging path.[7−10]

The phenomenological forces responsible for stabilizing the folded macromolecule and its association with other biomolecules usually include electrostatic interactions, covalent cross-linking, and hydrophobic interactions.[6] Many models of the stability for macromolecules have implicated effective hydrophobic interactions in the interior of the folded structure consisting mainly of constituents of varying hydrophobicity.[7,8] The effective interactions between individual hydrophobic moieties are relatively small, but the collective effects of these interactions and the consequences of desolvation in large regions in the interior or at an interface provide the stability seen in biomolecules containing a hydrophobic core. The degree to which the 3D-IEs predict these forces for different solution mixtures is somewhat ambiguous, and the study of more basic models is warranted to ascertain the qualitative predictions of the theory. These types of studies are needed for two reasons: to gain an understanding of the predictive quality of the present theory and to provide a path for relevant advancements to the theory.

A great deal of recent work has been applied to understanding the nature of hydrophobic effects at a fundamental

* Corresponding author e-mail: pettit@uh.edu.
[†] University of Houston.
[‡] Bhabha Atomic Research Centre.

10.1021/ct8002817 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/11/2008

level.[11−16] The nature of the shorter ranged attractions, less than a few molecular diameters, of hydrophobic species has been described with multiple mechanisms over the years.[17−23] Authors have attributed the nature of the hydrophobic interactions to arise from either a gain of entropy due to release of solvent structured at the hydrophobic interface,[22] disruption of the H-bonding network of the aqueous solvent,[21] the entropic cost of forming the cavity with respect to the size of the solvent,[19,20,23] or combinations of these. A more detailed discussion of these theories is given in the recent review by Ball.[17] Biochemically, hydrophobicity refers to a variable with a significant range and is not a property which is binary, i.e. on or off as defined by, for instance, a single contact angle criterion.

In addition to these shorter ranged interactions, a much longer less well understood interaction has been observed at length scales up to a few thousand molecular diameters.[24,25] Current explanations for such forces include the existence of microscopic bubbles causing a collapse of the intersolute region[24,26] and an electrostatic mechanism due to an induced dipole−dipole state from the slow reorientation of the polar solvent at the liquid−solid interface.[27,28] There are still differences in how different fields of science define hydrophobicity and the hydrophobic effect.[29] Here we concentrate on nanoscopic plates larger than atoms but still on the nanoscopic scale which have surface areas in the range of protein−protein contacts.

Simulations and theory have generated literature on a variety of solutes with varying hydrophobicity in polar solvents discussing the factors effecting the magnitude, the length scales of the interactions, and the time dependent kinetics.[15,16,30−32] For solutes in close proximity the factors contributing to the crossover from the solvent separated state to the contact states have been extensively studied.[15,18,33] The length scale of the crossover has been shown to be highly dependent on the solute−solvent attractive interactions.[15,33] For models with a repulsive or quite weak attractive interaction the intersolute region has been shown to go through a drying phase at larger distances of two or three solvent diameters.[33,34] The region between the solutes becomes depleted of the polar solvent and ultimately collapses through capillary evaporation. This intersolute collapse following dewetting, whether for hard sphere systems or for systems where dewetting is purely steric, is thermodynamically characterized by a large positive change in entropy due to the solvent being able to recoup some of the translation and rotational entropy due to the region being released to the bulk solvent.

Simulations of models having realistic solute−solvent attractive terms behave qualitatively differently than their purely repulsive models.[33−35] The state of hydration in the intersolute regions in simulations has been shown to be highly dependent on a number of factors characterized by the solute properties such as shape and interaction strength. For example, water was shown to exist in a stable one-dimensional hydrogen bonded chain in the interior of a carbon nanotube with sufficient attractive solute−solvent interactions but less so with lower attractions.[36,37] Other studies with small attractive interaction terms have shown a two-dimensional layer of water to be unstable when confined in a hydrophobic region at close contact.[15,38,39] A systematic analysis[33] used the calculated potential of mean force as a function of solute−solvent attraction and distance to determine the full range of hydrophobic effects on the association of 2-D sheets of atoms in aqueous solvent. Those authors showed the hydration state corresponding to the PMF at contact depends critically on the potential. They also observed that the stability of the monolayer of water between the plates at a separation consistent with the atomic diameters of the constituents of the plate-water-plate layers is strongly interaction potential strength dependent.

Simulations of solvation effects are considered the most accurate means to study such model systems at the molecular level, but they are computationally intensive. Once the solvent structure is determined via the pair distribution functions (PDF), the thermodynamic properties may be calculated. Many less expensive methods exist for such calculations, which vary in their qualitative and quantitative accuracy, but one of the most promising methods which still accounts for the atomistic nature of the solution species is based on integral equations.[4] Integral equations based on Ornstein−Zernike (OZ)-like constructs are attractive as an economical method for calculating the approximate pair distribution functions for the constituents of a solution.[40−42] The distribution functions predicted by the IEs can be used to describe much of the thermodynamics of the solution mixture but remain inherently approximate. Statistical thermodynamics provides convenient analytical expressions for the solution thermodynamics using the site−site distribution functions.[43,44] Some recent advances in IE theory offer more rigor and accuracy but have yet to be extended to aqueous systems.[41]

Extensions of the traditional one-dimensional (1D) radial methods to three-dimensional (3D) grid calculations show promise for the popular interaction site models[1−3,45] to obtain the solute molecule-solvent site pair distribution functions for a solute molecule at infinite dilution. The 3D-RISM theory, which is an extension of 1D-RISM theory,[46] requires the solvent−solvent correlations from a one-dimensional, radial theory and are used on a 3D grid to calculate the solute molecule-solvent site distributions. The distributions obtained from 3D-RISM provide an approximate angular distribution description of the solvent structure around solute molecules compared to the simple radial distribution functions (RDF) from 1D methods.

Some recent applications of the 3D-IEs have focused on elucidating the thermodynamic role the solvent plays on the conformational stability of large biomolecules such as proteins.[9,47,48] These studies considered the origin of the energetic and entropic contributions to the free energy and confirmed how the entropic part is a significant driving force in protein folding. An insightful discussion by those authors demonstrated the qualitative and quantitative nature of the results based on modern day solvation mechanisms. Other applications include the prediction of the spatial positioning of solvent species throughout possible cavities in the interior of large biomolecules, where cavities are not always accessible during MD simulation time scales, and the results may

be expected to show ensemble dependence.[49−51] In principle this is one of the advantages such methods have over molecular dynamics simulations outside of the grand canonical ensemble; they are able to sample all of configuration space, whereas NVE and NPT MD simulations can suffer from sampling errors due to time constraints.

The IEs and their closures are only approximations, and little has been discussed in the literature on the deficiencies of the 3D-IE theory as compared to volumes written on 1D theories compared with other methods (simulation or experiment). Comparison with the same Hamiltonian or potential model used in simulation for a nonpathological system eliminates the ambiguity of comparison with experiment. Once the accuracy of the theory for a given model is known, comparison with experiment can establish the precision of the model with confidence.

Here we address the question as to how accurately 3D-RISM equations describe a particular well studied system of interest, namely the 2D plates considered by this group earlier via simulation.[11,30,33,35] Some of the primary focus is on the dewetting transition seen in the simulation results and how the 3D-IEs coupled with the HNC closure will handle this transition. Expectations are limited since the HNC closure, like many other closures, is incapable of predicting coexisting phases.[52] This is mathematically shown in density functional theory to do with the second order truncation of the expansion in density of the thermodynamic potential and the reality that only a single minimum with respect to density occurs in a quadratic function.[52] The solvent behavior as described by IEs for small hydrophobic solutes has been studied in numerous papers.[12,14,53−55] The IE approach provides convenient analytical expressions for the solution thermodynamics using the site−site distribution functions.[43,44] The IEs in these studies generally provided qualitative descriptions of the solvation of small species. We also wish to quantitatively compare the effect of the solvent−solvent correlations from several IEs on the results for 2D plates. We will quantitatively compare the free energy from IE results for the state of hydration for various models of water near 2D plates of varying hydrophobicity with the exhaustive simulation data.

The IE approach is many orders of magnitude more computationally efficient than simulations, but questions about the accuracy and precision of the numerical solutions are still outstanding.[5] Numerical solutions for larger solutes which require a larger spatial grid can be difficult to obtain and can exhaust large amounts of memory. The convergence and memory issues are the dominating factors in considering which numerical method to use. Direct inversion of the iterative subspace (DIIS)[56−59] methods are known as being well balanced between memory requirements and convergence of the solutions. However the method is not an exact minimization routine. It is only as good as the solutions being used in its iterative subspace. Methods based on a Newton−Raphson type iterative routine are exact in that they will minimize the residual if the starting solution is within the radius of convergence. The negative aspect of these routines is dealing with a potentially large Jacobian matrix which has to be calculated, stored, and inverted.[60] In this paper we will present a method based on an exact iterative routine which does not require the storage of the full Jacobian or the costly matrix inversion.

In section II of this paper an introduction to the integral equations, the thermodynamic equations, and the numerical method used in this study are given. Section III describes the model interactions used in this study and in the corresponding simulation. In section IV the hydration structure and thermodynamics predicted by 3D-IEs is compared with simulation results to show the qualitative nature of the IE methods. Section V contains our conclusion.

## II. Theory and Methods

**A. Equations.** The three-dimensional molecular solute−solvent site distributions were calculated using the 3D-RISM IEs[1,3,45] for a multicomponent fluid using various approximations for the solvent distributions. For a solute molecule at infinite dilution the equation can be expressed as

$$h^{uv}(\mathbf{r}) = c^{uv}(\mathbf{r}) * (w^{vv}(\mathbf{r}) + \rho h^{vv}(\mathbf{r})) \quad (1)$$

where $h^{uv}(\mathbf{r})$ is the 3D solute−solvent pair total correlation function, $c^{uv}(\mathbf{r})$ is the 3D solute−solvent direct correlation function, $\rho$ is the particle number density of the solvent, $w^{vv}(\mathbf{r})$ is the intramolecular correlation function of the solvent, and $h^{vv}(\mathbf{r})$ is the solvent site−site distribution functions which we obtained from and will compare among various theories including RISM,[46] DRISM,[61,62] and closure optimized PISM variants.[40] The matrix convolution integral in real space is represented by an $*$. The hypernetted chain (HNC) equation was used in the 3D calculations as a closure. The HNC equation, although not exact, is computationally convenient and gives more reliable results for charged and polar species over a broader range of solution species than some other closure relations.[63] The HNC equation is defined for each component as

$$c_{ij}(r) = \exp(-\beta u_{ij}(r) + t_{ij}(r)) - t_{ij}(r) - 1 \quad (2)$$

where $t(r)$ is the indirect correlation function, $t(r) = h(r) - c(r)$, $u(r)$ is the pair potential between the molecular solute and solvent site, and $\beta$ is the inverse product of the solution temperature and Boltzmann's constant. We use this in the traditionally renormalized form.

The solvent site−site distributions in eq 1, $h^{vv}$, constitute a completely separate problem and are solved independently of the 3D calculations. The RISM equation for the solvent−solvent distributions is

$$h^{vv} = w^{vv} * c^{vv} * w^{vv} + w^{vv} * c^{vv} * \rho h^{vv} \quad (3)$$

where the correlation functions have the same meaning as in eq 1. Solutions based on XRISM theory are obtained from using eq 3 closed with the HNC (eq 2) in renormalized form.[64−66] For solution mixtures containing charged sites the correct screening due to the solvent is more accurately represented using dielectrically consistent RISM theory (DRISM).[61,67] DRISM theory is similar to RISM theory, but an ad hoc bridge term correcting the long-range behavior of the solvent is included in the closure. Methods based on RISM are not formally exact in that the graphical expansion

of the RISM equations includes incorrect terms and excludes other correct ones.[63] The proper interaction site method (PISM) is a formal improvement over RISM and more accurately represents the graphical expression but generally provides less accurate results.[68,69]

The hydration free energy (HFE) was calculated using the Morita-Hiroike HNC formula[43,44] on the 3D grid as

$$\Delta\mu = \rho k_B T \sum_a \int \left[\frac{1}{2}h_a(\mathbf{r})^2 - c_a(\mathbf{r}) - \frac{1}{2}h_a(\mathbf{r})c_a(\mathbf{r})\right]d\mathbf{r} \quad (4)$$

The chemical potential can be decomposed into its partial molar enthalpic, $\Delta\epsilon$, and entropic, $\Delta s$, parts using

$$\Delta\mu = \Delta\varepsilon^{uv} + \Delta\varepsilon^{vv} - T\Delta s \quad (5)$$

where the entropic part can be calculated by taking the temperature derivative of the chemical potential (eq 6).[53] The entropy is calculated using a finite difference method with a temperature step of 5 K.

$$-\frac{\partial\Delta\mu}{\partial T} = \Delta s \quad (6)$$

**B. Numerical Methods.** The correlation functions ultimately must be computed on a grid of sufficient resolution to capture the details of the finest physical length scales defined by the problem. With atomic diameters and bond lengths defined by angstroms, the grid size needs to be a fraction of an angstrom, and the extent of the grid must be well beyond the correlation lengths induced by the solute in the solvent. Given an appropriate grid, coarse solutions to eqs 1 and 2 could be calculated using a Newton−Raphson type numerical scheme on a smaller or reduced grid followed by Picard iterations on the full grid to refine the solution to the desired accuracy. Obtaining solutions in this manner however is usually not feasible due to memory constraints and the time required for large matrix inversions.[56] We have dealt with both of these issues to make these types of algorithms more tractable. The method presented here is an extension of the multigrid idea developed by Gillian[70] where a coarse solution is represented by a subset of points on a finer grid. The projection of the fine grid onto the coarse grid represents a significant reduction in the number of points used to describe the major features of the fluid. The coarse grid does not describe all the fine details in the solutions but can aid in the reduction of the number of Picard iterations needed on the fine grid to refine the solutions. On returning to the coarse grid values are calculated by averaging the functions at points in the fine grid.[60]

To reduce the computationally exhaustive task of inverting large matrices we used an iterative routine to approximate solutions for the linear equations. The algorithm we implemented for the solution to the linear equation, $Jx=F$, where $J$ is the Jacobian matrix, $x$ is our desired solution, and $F$ is the residual error, is the GMRes method.[71] Methods of this type, based on solving Newton's equations

$$0 = F(x) + \frac{\partial F(x)}{\partial x}\Delta x$$

$$\Delta x = \left(-\frac{\partial F(x)}{\partial x}\right)^{-1}$$

$$F(x) = (-J)^{-1}F(x)$$

are more stable than Picard or relaxation iterations and when implemented correctly accelerate the convergence of the solution.[72] The Generalized Minimal Residual (GMRes) routine iteratively forms an approximate solution to $x$ using $x_0 + z$, where $x_0$ is an initial starting guess and $z$ is the incremental solution in Krylov space. The approximating solution converges to the exact solution $x$ as the Krylov subspace approaches the rank of the transformation matrix $A$ representing the full set of equations. At each iteration a new approximation to $x$ is formed by minimizing the residual norm in the Krylov subspace, $K_k = \text{span}\ (v_1, Av_1, ..., A^{k-1}v_1)$.

$$\min_{z\in K_k} ||b - A[x_0 + z]|| = \min_{z\in K_k} ||r_0 - Az||$$

The $k^{\text{th}}$ iterative solution to $x$ using $z$ is of the form $x_k = x_0 + V_k y_k$, where $z = V_k y_k$. The matrix $V_k$ consists of $k$ columns, which are the $l_2$-orthonormal basis $(v_1, v_2, ..., v_k)$ of the Krylov subspace, and $y_k$ consists of the coefficients for the basis set. For each iteration the size of the Krylov subspace is increased by one, and the basis set coefficients are calculated to minimize the residual norm. This is continued until the residual norm meets some predetermined convergence criterion. The number of operations needed for each iteration is $O(n^2)$, where $n$ is the number of grid points. The number of iterations needed to converge to a solution depends on the initial basis set and the predetermined convergence criterion. An advantage to this routine is that the Jacobian matrix does not have to be stored, if one can efficiently generate the elements of the Jacobian (Appendix). However, the matrix consisting of the $l_2$-orthonormal basis should be stored and increases linearly in size with the number of iterations. A common variant of the GMRes routine is one that allows restarting. Once the number of basis vectors gets too large, the GMRes routine is restarted with the last approximate solution to $x$ as the new initial starting solution, $x_0$.

**C. Numerical Procedure.** After situating the solute coordinates in the approximate center of the 3-D grid, the steps used in our implementation are as follows:

1. Calculate the potential energy interactions between solute sites and solvent grid points.
2. Guess initial correlation functions.
3. If necessary perform a few initial Picard iterations.
4. Decompose correlation functions into their basis set coefficients and their fine displacements.
5. Calculate the residual vector.
6. Calculate the Jacobian elements on the coarse grid.
7. Perform GMRes routine until predetermined convergence criterion is met.
8. Calculate new basis coefficients.
9. Check residual vector to determine if another iteration on the coarse grid is needed (if so return to step 4) or to continue to fine grid (next step).
10. Calculate fine grid values from the basis set representation and the fine displacements.
11. Use Picard iterations to refine direct correlation, $c(\mathbf{r})$, solution in the fine grid.
12. Calculate fine grid residual vector $f^{i+1}(\mathbf{r})$. If this value meets our convergence criteria, the solution to the desired error has been obtained.

Numerical methods similar to the one presented here, which are based on the GMRes method, have previously been successfully applied to 1D site−site IEs.[72,73] A final thought on these types of numerical routines is that the responsibility of calculating the solution to the desired criterion is up to the fine grid iterations. For solutions to nonstiff problems a coarser grid can be used to decrease the amount of time needed per iteration, and for stiff problems where convergence issues are the main concern more points should be used in the coarse grid.

## III. Model

In this study the integral equations described above are used to study the effective interactions between two planar hydrophobic solutes in a polar solvent of $H_2O$. The results are then compared with simulation results for the same model.[33] The solute and solvent parameters are identical to the simulation study. The pair potential between the sites is modeled with the Lennard-Jones (LJ) plus Coulomb potential. The potential for the solute−solvent interactions in this study is calculated using

$$u_{ab}(r) = 4\varepsilon_{ab}\left[\left(\frac{\sigma_{ab}}{r}\right)^{12} - \left(\frac{\sigma_{ab}}{r}\right)^{6}\right] + \frac{q_a q_b}{r} \qquad (7)$$

The subscripts specify the species of the solvent defined on a grid for the 3D calculations. The symbols $\sigma$, $\epsilon$, and $q$ are the usual LJ plus Coulomb parameters: the diameter, the well depth, and the charge on the atomic sites, respectively. The $H_2O$ molecules are represented by the SPC/E model.[74] The LJ parameters for this SPC/E $H_2O$ model have the values of $\sigma_{OO} = 3.1655$ Å, $\epsilon_{OO} = 0.1554$ kcal mol$^{-1}$, and $q_O = -0.8476$ for the sites characterizing the oxygen atom and $\sigma_{HH} = 0.4$ Å, $\epsilon_{HH} = 0.02$ kcal mol$^{-1}$, and $q_H = 0.4238$ for the sites characterizing the hydrogen atoms. The attractive part of the Lennard-Jones potential is excluded for interactions involving the solvent hydrogen atoms as is consistent with previous usage. The two molecular plates are represented by flat graphene sheets of carbon atoms, the same used by Choudhury and Pettitt.[33] The sheets consisted of 60 carbon atom sites each and measured ~11 × 12 Å. The carbon atoms have been placed in a rigid hexagonal graphite-like pattern with distances of 1.4 Å between nearest neighbors. The LJ parameters for the carbon atoms are $\sigma_{CC} = 3.4$ Å and $\epsilon_{CC} = 0.3598$ kJ mol$^{-1}$, and the sites are uncharged. These parameters represent sp$^2$ carbon atoms from the AMBER 96 force field from the hydrophobic aromatic peptide residues. The Lorentz−Berthelot combination rules, $\sigma_{OC} = (\sigma_{OO} + \sigma_{CC})/2$ and $\epsilon_{OC} = (\epsilon_{OO}\,\epsilon_{CC})^{1/2}$, are used to calculate the LJ parameters for site−site interactions between different species. For some calculations the solute−solvent attractions are turned off by using this same potential with a Week-Chandler-Anderson decomposition.[5] The number density of the bulk $H_2O$ was 0.03334 Å$^{-3}$, and the temperature is 298.15 K.

The site−site radial distributions for the $H_2O$ solvent, $h^{\nu\nu}$, are calculated using XRISM, DRISM, and optimized PISM theories.[40,61,66,67] The results between XRISM and DRISM are quantitatively indistinguishable for the charge free solutes used here (see Figure 1), so only the distributions from

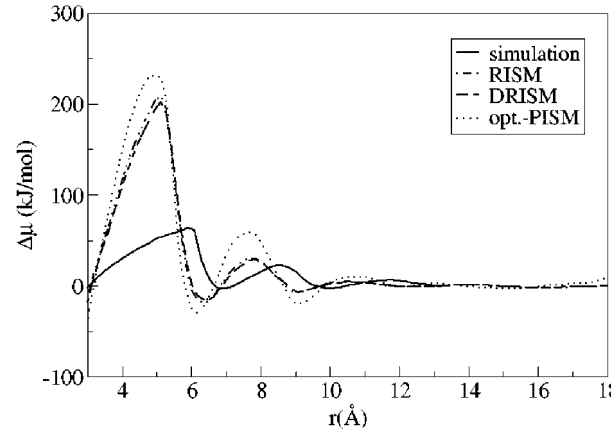



***Figure 1.*** Solvent contributions (HFE) to the PMF for different $h^{\nu\nu}$ $H_2O$ theories.

DRISM theory are shown in the results after Figure 1. The radial solvent−solvent distributions are calculated on a fine grid of 16,364 points over a range of 40 Å ($\Delta r = 0.00244$ Å) to minimize errors when transferred to the 3D grids. To limit the errors in the authenticity of the solution and the thermodynamic variables calculated from 3D distributions the calculations are routinely converged to a relative residual error of $10^{-12}$. The 3D calculations are performed in a cubic box measuring 40 Å to a side. The fine grid is defined by a linear grid consisting of $128^3$ points, and the coarse grid consisted of $19^3$ points. We note that some poor initial guesses for thermodynamic values not in a stable part of the phase diagram result in solutions that initially appear to start to converge down to a relative tolerance of $10^{-3}$ or $10^{-4}$. These solutions then diverge on further attempts to refine them and so are rejected in our work as unphysical unless stable solutions at our most stringent criteria are met.

## IV. Results and Discussion

The 3D distribution functions for the parallel graphene plates in solvent $H_2O$ were calculated for intersolute distances ranging from 1.0−18.4 Å in increments of 0.1 Å. The HNC hydration free energies (HFE) from the 3D-RISM results shown in Figure 1 were calculated using the chemical potential formula[44] from eq 4. The value for the infinitely separated plates was subtracted from each curve for each solvent theory. The sensitivity of RISM-like absolute free energies or chemical potentials is well-known.

The three sets of data in Figure 1 correspond to the different solvent−solvent distributions obtained from XRISM, DRISM, and optimized PISM theory. All of the IE results show qualitative features but are quantitatively disappointing in both the magnitude and phase of the oscillations. As mentioned above, the RISM and DRISM results are similar since graphite has no atomic site charges, so only the DRISM results are displayed throughout the remainder of the analysis. Optimized PISM theory is a recent method for calculating the distribution functions which compares to simulation better than DRISM for pure solvents.[40] Optimized PISM theory has distributions which are more consistent with the simulation data for the solvent−solvent correlations. However, in this usage it appears to over-represent the free energy

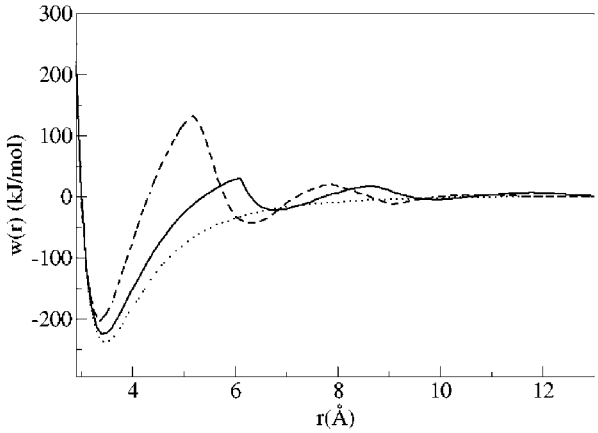



***Figure 2.*** Potential of mean force for DRISM (dashed line) and MD (solid line) results plotted with the direct potential (dotted line) interaction between the plates.

contributions or features from the solvent when compared to the simulation and DRISM results. It should be mentioned that no separate solute−solvent closure optimization was done here as would be more consistent with that method.[40]

The HFE contains the contributions to the PMF from direct solvent interactions. The PMF obtained by the addition of the HFE and the direct solute−solute potential is shown in Figure 2 for the DRISM solvent.

The minimum in the potential of mean force for the solutes in the solute contact configuration (no intervening water layers) predicted by the 3D-RISM theory is 3.4 Å. This distance is ∼0.1 Å less than the distance at which the minimum occurred in the PMF obtained from the MD simulation study.[33] Both distances are slightly less than the minimum shown in the direct potential. An explanation for these slightly shorter distances is due to the slope of the cavity potential as the plates are forced together which has been seen in RISM style calculations before.[75] The generally incorrect placement of the HFE to the direct potential for RISM-like theories has been noted previously and has a variety of consequences for the resulting PMF and properties.[75]

The PMF shown in Figure 2 for the DRISM solvent has three significant solvent stabilized minima occurring at distances of 6.5, 9.3, and 12.5 Å. These solvent stabilized minima occur at distances allowing well defined solvent layers between the plates as shown in Figure 3 and expected from simulation. The first, second, and third solvent separated minima occur at configurations allowing one, two, and three intervening water layers between the plates, respectively.

The same characteristics are seen in the results from the simulation study.[33] However, the simulation results predicted these minima at slightly larger distances consistent with the phase shift in the PMF noted earlier. The DRISM and MD results[33] for the solute−solvent oxygen site distributions perpendicular to the plate surfaces for the three minima corresponding to the distances at which the minima occur in the simulations are shown in Figure 4.

The distributions shown in Figure 4 from DRISM are structurally similar to the distributions obtained from MD simulations with some noted differences. For the intersolute

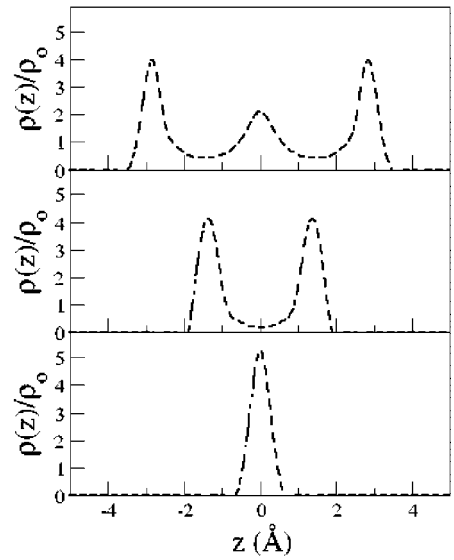



***Figure 3.*** Solute−solvent (oxygen) distributions at the solvent stabilized solute configurations for DRISM at plate separations of 6.5 (bottom), 9.3 (middle), and 12.5 (top).

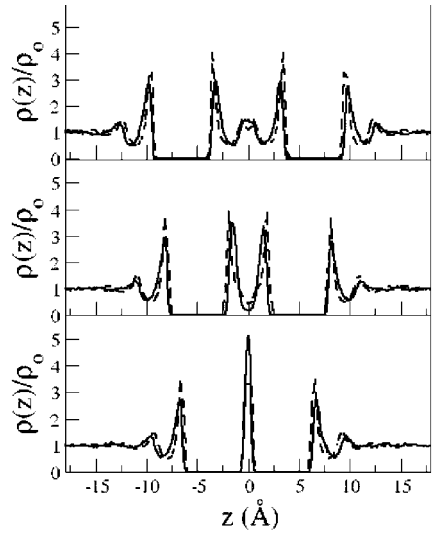



***Figure 4.*** Oxygen site distributions for simulation and DRISM at distances of 6.8 (bottom), 9.8 (middle), and 13 (top), corresponding to the solvent stabilized minima observed in the PMF from simulation.

distances displayed in Figure 4 all theories predicted well defined peaks for the solvent in contact with the outer surface of the plates. Comparisons with simulation are most favorably made when considering the predicted PMF minima; however, the minima are not found at identical distances.

The solvent distributions at the intersolute distance corresponding to the first solvent separated minimum are characterized by a large single peak between the plates. This represents a well ordered monolayer between the plates. However, this minimum in the PMF for the MD simulation occurs at 6.8 Å which is larger than the DRISM results by about 0.3 Å. The distributions for the second solvent separated minimum are also qualitatively similar between the two methods and represent an ordered bilayer of water between the plates. The distance between the peaks in the solvent distributions between the plates is 2.9 Å, which

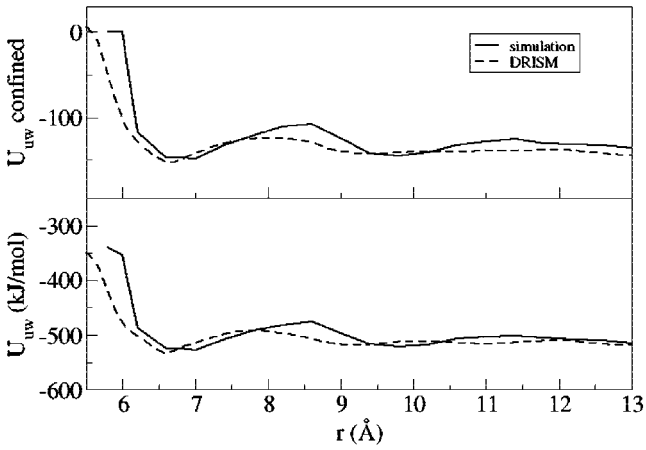



***Figure 5.*** The solute−solvent total potential energy (bottom) and the potential energy between the plates and solvent in the intersolute region (top) in kJ mol$^{-1}$ for IEs and simulation.

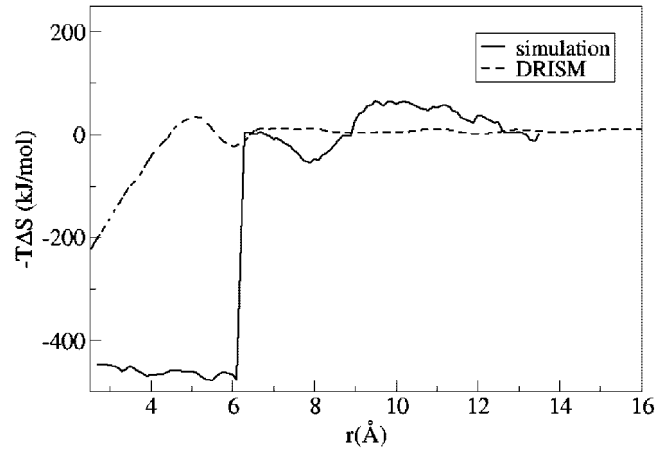



***Figure 6.*** Entropic contribution to the HFE ($-T\Delta S$).

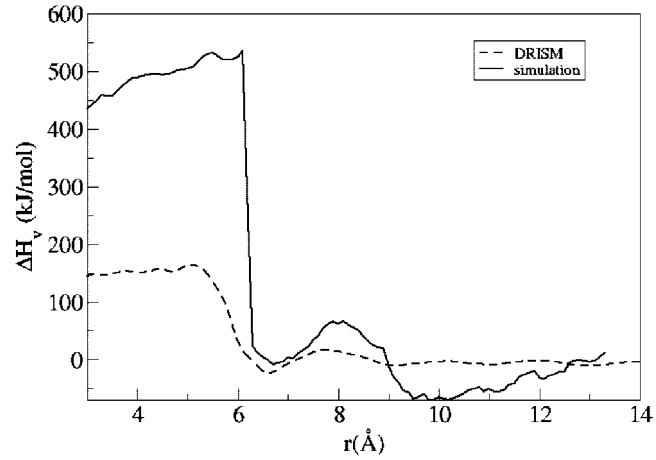



***Figure 7.*** Enthalpic contributions to the HFE, $H^v = \Delta\varepsilon^{uv} + \Delta\varepsilon^{vv}$.

roughly corresponds to the peak for O−O distributions of water. However, the interplate distance of this feature in the DRISM calculations is about 0.5 Å less than the plate distance from the MD simulation.

There is a third solvent stabilized configuration corresponding to three layers of water between the plates which is also replicated by the IE method. The inconsistency in the locations of the solvent separated minima between IEs and simulation are more pronounced and appear to be due to differences in the effective widths of the solvent structure layers. The structure of the solvent layers for the configurations corresponding to the free energy barriers—the regions between the solvent separated minima— in the PMF are characterized by the expected zigzag layering of water between the plates, which is also seen in the MD simulations.[76] The consequence of this layering is a loss of solute−solvent energetic interactions at each surface facing the interior region.

The most prominent characteristic of the PMF from the IEs is the large free energy barrier between the contact state and the first solvent separated state. The IE approach estimates the magnitude of this free energy barrier to be significantly larger than the barrier seen in simulations. The other obvious feature observed in the PMF for both the theories and simulation is a cusp that occurs at the start of the steric drying distance. The solute−solvent interaction energy helps explain the changes occurring in the free energy (Figure 2).

The qualitative energetic features of the solute−solvent interactions in the interior region are reproduced by the 3D-IE theory when compared to MD simulations. As the interplate distance decreases from the first solvent separated state the rise in the PMF is mainly attributed to the loss of solute−solvent interactions in the interior region. The cusp in the PMF for MD simulations is sharper because of the well defined, steric-induced loss of these favorable energetic interactions as the monolayer of water is effectively excluded from the interior region. This feature can be seen in Figures 2 and 5 at ∼6.0 Å. The cusp is not as sharp in the PMF from the IE theory because the IEs allow some probability of water molecules to remain in sterically forbidden positions located above the center of the hexagonal solute atom configuration (see below).

To better understand the factors contributing to the HFE and to investigate the reason for the large free energy barrier in the PMF from IEs, the HFE was separated into its enthalpic and entropic contributions. The partial molar entropy and enthalpy are shown in Figures 6 and 7.

The nature of the larger free energy barrier is obvious when comparing the factors contributing to the PMF from simulation and IEs. Both methods predict a large positive change in solute−solvent interactions as the solvent is forced from the intersolute region upon decreasing the plate distance. However, in the simulation result there is a substantial favorable response in the entropy of the solvent which counteracts the energy lost in the solute−solvent interactions. This rapid change in entropy is not seen in the IE results; there is nonetheless a gradual gain in entropy as the intersolute distance decreases, which is most likely attributed to the gain in translational and rotation entropy as the excluded volume regions overlap or shrink.

As stated earlier, one of the more difficult features of hydrophobic interactions is the length scale at which drying between the plates occurs. In the MD simulations with full solute−solvent interactions[33] the smallest interplate distance

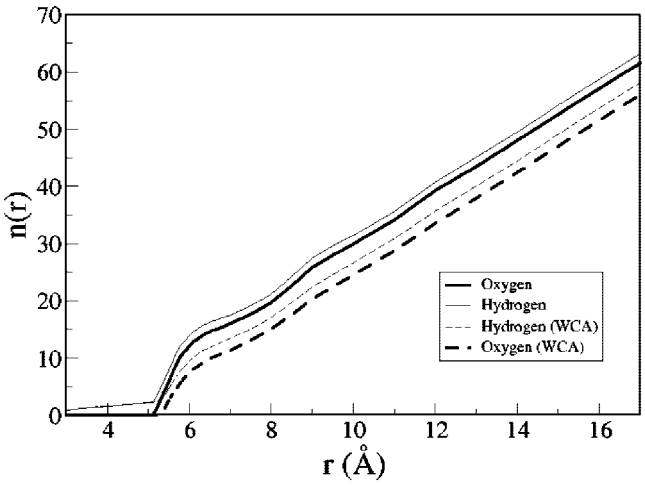



***Figure 8.*** Water oxygen and the sum of hydrogen solvent sites between the plates from DRISM for the model presented and a model represented with the WCA interaction potential.

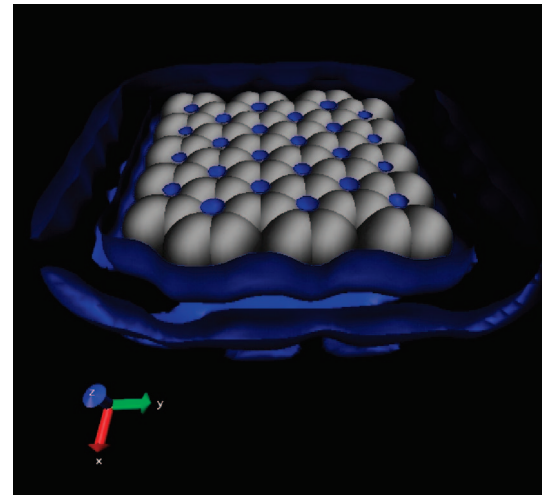


***Figure 9.*** Water hydrogen sites between the plates at a separation of 3.0 Å. Shows the existence of finite probabilities of hydrogen between the plates when no oxygens are present, i.e. in carbon−carbon contact.

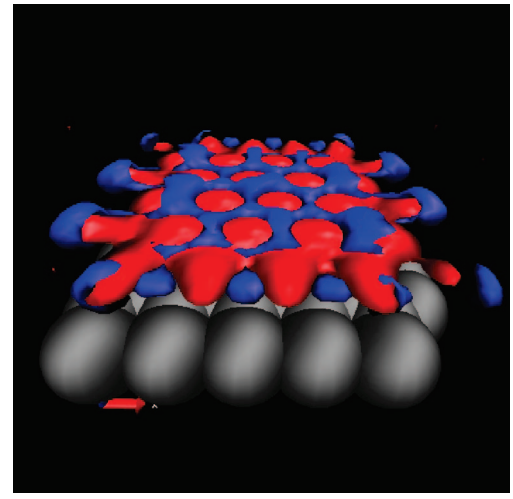


***Figure 10.*** Isosurface of the water oxygen (red) and hydrogen (blue) sites between the plates at a separation of 6.8 Å.

at which water was allowed between the plates is 6.2 Å; any shorter a distance and all the water molecules were excluded due to steric effects. The sharp change in the number of water molecules between the plates was not duplicated by the 3D-IEs with any solvent theory used here. An observation of the distribution functions at these shorter distances shows that there is some nontrivial probability for the water molecules to be located in the sterically forbidden region. Although not explicitly observed by the number of water molecules between the plates, there is evidence that the IEs have a rough qualitative sense of the beginning of the steric drying transition as seen in the MD simulations. For the interplate region to be hydrated by water molecules in a hydrogen bonded network forming a two-dimensional sheet ~13 water molecules would be needed to cover the area. The distance at which the number of water molecules falls below 13 in the IEs is about 6.1 Å, as shown in Figure 8, which is near the same length of steric drying in the MD simulations. It is however, unphysically gradual, and so the cavity contains significant density when just steric effects would have it be dry.

The number of solvent sites between the plates for a model with no solvent−solute attraction (via the WCA decomposition)[77] is also shown in Figure 8. Here we expect to demonstrate the drying effects due to the lack of attraction of the plates. In contradiction to the finding from simulations, the lack of attraction with the IEs does not lead to drying starting at over 10 Å as has been seen and verified by simulation.[33,34] In fact the dramatic effect of solute−solvent attractions seen in simulations is effectively suppressed. The lack of a transitional state for the coexistence of phases has been noted before for integral equation methods.[52] This brings into question the reliability of IEs in situations where a molecular sized cavity might not have sufficient interactions to stabilize a water molecule e.g. on the hydrophobic interior of a protein.

Figure 8 also shows the inconsistency in the stoichiometry of the water oxygen and hydrogen sites expected from all XRISM-type theories.[63] Only half the hydrogen site densities are shown so the ratio should be 1:1, but there are consistently more hydrogen atoms between the plates. This behavior continues past the point where all water oxygen sites are excluded from the intersolute region—the number of hydrogen sites observed should then be zero by stoichiometry. Figure 9 shows the location of the nonzero probability of these rogue hydrogen atoms between the plates. Note that no dissociation is possible for the model, and thus this is a pure artifact of the methods.

This stoichiometric problem has been noted for XRISM class theories for some time.[78] Another property of the solvent in the interplate region observed in the MD simulation, which is also computable in the 3D-IE theories, is the orientation of water and the static hydrogen bonding occurring between the plates with a single water layer. The isosurface plot shown in Figure 10 shows the oxygen atoms (red) tend to occupy the positions directly above the center of the hexagonal pattern. There is both a steric hole as well as an attractive maximum in the potential in these positions.

The hydrogen atoms (blue) tend to occupy the positions between the oxygen atoms forming a hydrogen bonding network. The overall orientation of the water molecules is for the plane through HOH to be just above or below the surface of the oxygens which is the same orientational preference observed in MD simulations and in good accord.

## V. Conclusion

In this study we used the 3D-RISM IEs with XRISM, DRISM, and optimized PISM solvent−solvent distributions to calculate the free energies and distribution functions of the solvent water sites around large graphene plates. The graphene sheets were modeled by Lennard-Jones potentials from Amber describing $sp^2$ carbon atom sites with and without the WCA decomposition. The SPC/E water model was used for the solvent. The PMF as a function of interplate distance was calculated to determine the ability of the IEs to predict the physical behavior of the solute/solvent combination types presented. The thermodynamic values and structural properties obtained by the IEs were compared to the same from MD simulation results for the same model. The gross structural details of the solvent in contact with the hydrophobic surfaces were consistent between the theories and in reasonable accord with simulation with some notable exceptions. The structural layering of water between the sheets seen in the 3D-IEs was somewhat shifted in position and included stoichiometric deviations compared to the MD simulations.

The free energy barrier for the transition from the first solvent separated configuration to contact was qualitatively replicated by all the IE theories but was off quantitatively. This barrier was significantly larger for the IE results compared to the MD simulation results. The barrier was shown to be dependent on the expected translation-rotation entropy effects for solvent release (increasing intersolute distance) and solute−solvent interaction energies for association (decreasing intersolute distance). The difference in the magnitude of the first free energy barrier from IEs and simulations was shown to arise from the inability of the IEs to accurately account for changes in the solvent entropy and the compensating solvent enthalpy when the solvent was excluded from the interplate region.

The abrupt change in the number of water molecules between the plates at the transition distance for simulation was not seen by the IE theory and may also reflect the standing controversy in the two-state versus glassy debate for such systems. However, the distance at which the number of water molecules between the plates predicted by the IE theory would not completely hydrate the surface was similar to the distance at which the MD simulations saw dramatic dewetting in the intersolute region. The IEs were also shown to predict a reasonable orientation of water molecules in contact with the plate surfaces as compared with the MD simulations.

The calculations at high precision shown in this work were done with a combination of Picard/Newton−Raphson based method to obtain the solutions to the 3D-IEs. We showed how the problem of calculating and storing a large Jacobian can be replaced with a method to generate the analytically calculated Jacobian elements as they are needed for a smaller system of equations. Although this numerical method will not completely replace DIIS methods, it can help to converge solutions which appear to be unstable with DIIS, especially in the early iterations.

The accuracy of the probability of water in confined spaces and on the interior of macromolecules predicted by this class of theories may suffer from problems found with our more ideal systems. The lack of sharp (first order) hydrophobic drying and inconsistent stoichiometry are problematic. Grand canonical simulations would more accurately quantify this in biomolecule interiors where entrance and escape by diffusion can be problematic.

**Acknowledgment.** Dr. Marcelo Marucho is thanked for the optimized PISM results and many stimulating conversations. This work was supported by the National Institutes of Health (GM066813), the Robert A. Welch Foundation (E-1028), and a training fellowship to J.J.H. from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM grant No. 5T15LM07093). B.M.P. thanks Professor Fumio Hirata for a thoughtful discussion on this manuscript.

## Appendix

To calculate the Jacobian elements we use the following equation, where the second equation is shown in Fourier space and in matrix form.

$$F(c(r)) = \exp(-\beta u(r) + t(r)) - t(r) - c(r) - 1 \quad \text{(A1)}$$

$$\hat{h}(k) = \hat{c}(k)\hat{w}(k) + \hat{c}(k)p^{v}\hat{h}(k) \quad \text{(A2)}$$

The elements of the Jacobian matrix in this implementation are defined as

$$J^{a,i}_{b,j} = \frac{\partial F^{ua}(c^{ua}(r_i))}{\partial c^{ub}(r'_j)} = -\delta_{(i,j)}\delta_{(a,b)} + \frac{\partial F^{ua}(c^{ua}(r_i))}{\partial t^{ua}(r_i)}\frac{\partial t^{ua}(r_i)}{\partial c^{ub}(r'_j)} \quad \text{(A3)}$$

where the superscripts denote the functions between the solute and the solvent sites, and the subscripts on $r$ denote the position in coordinate space.

The partial derivatives can be calculated using the site−site representation of the HNC closure and the OZ equation in real space. The partial derivative can be calculated as

$$\begin{aligned}\frac{\partial F^{ua}(c^{ua}(r_i))}{\partial t^{ua}(r_i)} &= \exp(-\beta u^{ua}(r_i) + t^{ua}(r_i)) - 1 \\ &= h^{ua}(r_i)\end{aligned} \quad \text{(A4)}$$

To calculate the second partial derivative in the last term we start with the OZ equation in momentum space, A.2, and expand into the components for all constituents

$$\begin{aligned}\hat{h}^{ua}(k) &= \sum_b \hat{c}^{ub}(k)\hat{w}^{ba}(k) + p^{v}\sum_b \hat{c}^{ub}(k)\hat{h}^{ba}(k) \\ \hat{h}^{ua}(k) &= \sum_b \hat{c}^{ub}(k)(\hat{w}^{ba}(k) + p^{v}\hat{h}^{ba}(k))\end{aligned} \quad \text{(A5)}$$

Next we can rearrange and collect the constant terms into $\chi^{ba}$, which is calculated in Fourier space

$$\hat{t}^{ua}(k) = \hat{c}^{ub}(k)p^v\hat{h}^{aa}(k) + \sum_{b\neq a}\hat{c}^{ub}(k)(\hat{w}^{ba}(k) + p^v\hat{h}^{ba}(k)) \tag{A6}$$

$$\hat{t}^{ua}(k) = \hat{c}^{ub}(k)p^v\hat{h}^{aa}(k) + \sum_{b\neq a}\hat{c}^{ub}(k)\hat{x}^{ba}(k) \tag{A7}$$

This equation maybe transformed back into real space to give

$$t^{ua}(\vec{r}) = \rho^v\int_o^\infty c^{ua}(\vec{r'}) * h^{aa}(\vec{r}-\vec{r'})d\vec{r'} + \sum_{b\neq a}\int_o^\infty c^{ua}(\vec{r'}) * x^{ua}(\vec{r}-\vec{r'})d\vec{r'} \tag{A8}$$

where the integrals run over all space.

The equation on a discretized grid in Cartesian coordinates, letting $r = (x,y,z)$ and $r' = (x',y',z')$ can be recast as

$$t^{ua}(x,y,z) = \rho^v\sum_{x',y',z'}c^{ua}(x',y',z')*h^{aa}(x-x',y-y',z-z')\Delta x'\Delta y'\Delta z' + \sum_{b\neq a}\sum_{x',y',z'}c^{ub}(x',y',z')*\chi^{ba}(x-x',y-y',z-z')\Delta x'\Delta y'\Delta z' \tag{A9}$$

$$\frac{t^{ua}(x,y,z)}{c^{ua}(x',y',z')} = \rho^v h^{aa}(x-x',y-y',z-z')\Delta x'\Delta y'\Delta z' \tag{A10}$$

$$\frac{t^{ua}(x,y,z)}{c^{ub}(x',y',z')} = \chi^{ba}(x-x',y-y',z-z')\Delta x'\Delta y'\Delta z' \tag{A11}$$

for $a = b$, where $i = (x,y,z)$ and $j'=(x',y',z')$

$$J^{a,i}_{a,j'} = \frac{\partial F^{ua}(c^{ua}(i))}{\partial c^{ub}(j')} = -\delta_{(i,j')} + (\exp(-\beta u^{ua}(i) + t^{ua}(i)) - 1)\rho^v h^{aa}(i\text{-}j')\Delta x'\Delta y'\Delta z' \tag{A12}$$

$$= -\delta_{(i,j')} + h^{ua}(i)\rho^v h^{aa}(i-j')\Delta x'\Delta y'\Delta z' \tag{A13}$$

and for $a \neq b$

$$J^{a,i}_{b,j'} = \frac{\partial F^{ua}(c^{ua}(i))}{\partial c^{ub}(j')} = (\exp(-\beta u^{ua}(i) + t^{ua}(i)) - 1)\chi^{ab}(i\text{-}j')\Delta x'\Delta y'\Delta z' \tag{A14}$$

$$= h^{ub}(i)\chi^{ab}(i-j')\Delta x'\Delta y'\Delta z' \tag{A15}$$

## References


(1) Beglov, D.; Roux, B. Numerical solution of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions. *J. Chem. Phys.* **1995**, *103* (1), 360–4.

(2) Beglov, D.; Roux, B. Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B* **1997**, *101* (39), 7821–7826.

(3) Cortis, C. M.; Rossky, P. J.; Friesner, R. A. A three-dimensional reduction of the Ornstein-Zernike equation for molecular liquids. *J. Chem. Phys.* **1997**, *107* (16), 6400–6414.

(4) Hirata, F. *Molecular Theory of Solvation*; Kluwer Academic Publishers: Dordrecht, 2003; Vol. 24.

(5) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic Press: San Diego, 1986.

(6) Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J.* **1996**, *10* (1), 75–83.

(7) Miranker, A. D.; Dobson, C. M. Collapse and cooperativity in protein folding. *Curr. Opin. Struct. Biol.* **1996**, *6* (1), 31–42.

(8) Itzhaki, L. S.; Evans, P. A.; Dobson, C. M.; Radford, S. E. Tertiary Interactions in the Folding Pathway of Hen Lysozyme: Kinetic Studies Using Fluorescent Probes. *Biochemistry* **1994**, *33* (17), 5212–20.

(9) Imai, T.; Harano, Y.; Kinoshita, M.; Kovalenko, A.; Hirata, F. Theoretical analysis on changes in thermodynamic quantities upon protein folding: Essential role of hydration. *J. Chem. Phys.* **2007**, *126* (22), 225102/1–225102/9.

(10) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Current Opinion in Structural Biology* **2004**, *14* (1), 70–75.

(11) Choudhury, N.; Pettitt, B. M. The Dewetting Transition and The Hydrophobic Effect. *J. Am. Chem. Soc.* **2007**, *129* (15), 4847–4852.

(12) Lum, K.; Chandler, D.; Weeks, J. D. Hydrophobicity at Small and Large Length Scales. *J. Phys. Chem. B* **1999**, *103* (22), 4570–4577.

(13) Pratt, L. R.; Pohorille, A. Hydrophobic Effects and Modeling of Biophysical Aqueous Solution Interfaces. *Chem. Rev.* **2002**, *102* (8), 2671–2691.

(14) Moghaddam, M. S.; Chan, H. S. Pressure and temperature dependence of hydrophobic hydration: Volumetric, compressibility, and thermodynamic signatures. *J. Chem. Phys.* **2007**, *126* (11), 114507/1–114507/15.

(15) Giovambattista, N.; Rossky, P. J.; Debenedetti, P. G. Effect of pressure on the phase behavior and structure of water confined between nanoscale hydrophobic and hydrophilic plates. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2006**, *73* (4−1), 041604/1−041604/14.

(16) Huang, D. M.; Chandler, D. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (15), 8324–8327.

(17) Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **2008**, *108* (1), 74–108.

(18) Pratt, L. R.; Chandler, D. Theory of the hydrophobic effect. *J. Chem. Phys.* **1977**, *67* (8), 3683–704.

(19) Lee, B. Solvent reorganization contribution to the transfer thermodynamics of small nonpolar molecules. *Biopolymers* **1991**, *31* (8), 993–1008.

(20) Lee, B. The physical origin of the low solubility of nonpolar solutes in water. *Biopolymers* **1985**, *24* (5), 813–23.

(21) Blokzijl, W.; Engberts, J. B. F. N. Hydrophobic effects: opinion and fact. *Angew. Chem.* **1993**, *105* (11), 1610–48. See also *Angew Chem., Int. Ed. Engl.* **1993**, *32*(11), 1545−79.

(22) Frank, H. S.; Evans, M. W. Free volume and entropy in condensed systems. III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. *J. Chem. Phys.* **1945**, *13*, 507–32.

(23) Lucas, M. Size effect in transfer of nonpolar solutes from gas or solvent to another solvent with a view on hydrophobic behavior. *J. Phys. Chem.* **1976**, *80* (4), 359–62.

(24) Parker, J. L.; Claesson, P. M.; Attard, P. Bubbles, cavities, and the long-ranged attraction between hydrophobic surfaces. *J. Phys. Chem.* **1994**, *98* (34), 8468–80.

(25) Wood, J.; Sharma, R. How Long Is the Long-Range Hydrophobic Attraction. *Langmuir* **1995**, *11* (12), 4797–802.

(26) Attard, P. Bridging Bubbles between Hydrophobic Surfaces. *Langmuir* **1996**, *12* (6), 1693–5.

(27) Despa, F.; Berry, R. S. The origin of long-range attraction between hydrophobes in water. *Biophys. J.* **2007**, *92* (2), 373–378.

(28) Despa, F.; Fernandez, A.; Berry, R. S. Dielectric Modulation of Biological Water. *Phys. Rev. Lett.* **2004**, *93* (22), 228104/1–228104/4.

(29) Maibaum, L.; Chandler, D. Segue between Favorable and Unfavorable Solvation. *J. Phys. Chem. B* **2007**, *111* (30), 9025–9030.

(30) Choudhury, N.; Pettitt, B. M. Dynamics of Water Trapped between Hydrophobic Solutes. *J. Phys. Chem. B* **2005**, *109* (13), 6422–6429.

(31) Pratt, L. R. Molecular theory of hydrophobic effects: "she is too mean to have her name repeated". *Annu. Rev. Phys. Chem.* **2002**, *53*, 409–436.

(32) Walther, J. H.; Jaffe, R.; Halicioglu, T.; Koumoutsakos, P. Carbon nanotubes in water: Structural characteristics and energetics. *J. Phys. Chem. B* **2001**, *105* (41), 9980–9987.

(33) Choudhury, N.; Pettitt, B. M. On the Mechanism of Hydrophobic Association of Nanoscopic Solutes. *J. Am. Chem. Soc.* **2005**, *127* (10), 3556–3567.

(34) Huang, X.; Margulis, C. J.; Berne, B. J. Dewetting-induced collapse of hydrophobic particles. *Proc. Natl. Acad. Sci.* **2003**, *100* (21), 11953–11958.

(35) Choudhury, N.; Pettitt, B. M. Enthalpy-Entropy Contributions to the Potential of Mean Force of Nanoscopic Hydrophobic Solutes. *J. Phys. Chem. B* **2006**, *110* (16), 8459–8463.

(36) Hummer, G.; Rasalah, J. C.; Noworyta, J. P. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature (London)* **2001**, *414* (6860), 188–190.

(37) Sansom, M. S. P.; Biggin, P. C. Biophysics: Water at the nanoscale. *Nature (London)* **2001**, *414* (6860), 156, 157–159.

(38) Wallqvist, A.; Berne, B. J. Computer Simulation of Hydrophobic Hydration Forces on Stacked Plates at Short Range. *J. Phys. Chem.* **1995**, *99* (9), 2893–9.

(39) Wallqvist, A.; Gallicchio, E.; Levy, R. M. A Model for Studying Drying at Hydrophobic Interfaces: Structural and Thermodynamic Properties. *J. Phys. Chem. B* **2001**, *105* (28), 6745–6753.

(40) Marucho, M.; Montgomery Pettitt, B. Optimized theory for simple and molecular fluids. *J. Chem. Phys.* **2007**, *126* (12), 124107/1–124107/9.

(41) Dyer, K. M.; Perkyns, J. S.; Pettitt, B. M. A site-renormalized molecular fluid theory. *J. Chem. Phys.* **2007**, *127* (19), 194506/1–194506/14.

(42) Dyer, K. M.; Perkyns, J. S.; Pettitt, B. M. Effective density terms in proper integral equations. *J. Chem. Phys.* **2005**, *123* (20), 204512/1–204512/11.

(43) Morita, T., III *Prog. Theor. Phys.* **1961**, *25* (4), 537–578.

(44) Singer, S. J.; Chandler, D. Free energy functions in the extended RISM approximation. *Mol. Phys.* **1985**, *55* (3), 621–5.

(45) Kovalenko, A.; Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chem. Phys. Lett.* **1998**, *290* (1,2,3), 237–244.

(46) Chandler, D.; ersen, H. C. Optimized cluster expansions for classical fluids. II. Theory of molecular liquids. *J. Chem. Phys.* **1972**, *57* (5), 1930–7.

(47) Phongphanphanee, S.; Yoshida, N.; Hirata, F. On the Proton Exclusion of Aquaporins: A Statistical Mechanics Study. *J. Am. Chem. Soc.* **2008**, *130* (5), 1540–1541.

(48) Imai, T.; Harano, Y.; Kinoshita, M.; Kovalenko, A.; Hirata, F. A theoretical analysis on hydration thermodynamics of proteins. *J. Chem. Phys.* **2006**, *125* (2), 024911/1–024911/7.

(49) Yoshida, N.; Phongphanphanee, S.; Maruyama, Y.; Imai, T.; Hirata, F. Selective Ion-Binding by Protein Probed with the 3D-RISM Theory. *J. Am. Chem. Soc.* **2006**, *128* (37), 12042–12043.

(50) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Locating Missing Water Molecules in Protein Cavities by the Three-Dimensional Reference Interaction Site Model Theory of Molecular Solvation. *Wiley InterScience* **2006**, *66*, 804–813.

(51) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Water Molecules in a Protein Cavity Detected by a Statistical-Mechanical Theory. *J. Am. Chem. Soc.* **2005**, *127* (44), 15334–15335.

(52) Evans, R., Density functionals in the Theory of Nonuniform Fluids. In *Fundamentals of Inhomogenous Fluids*; Henderson, D., Ed.; Marcel Dekker, Inc.: New York, 1992; pp 85−176.

(53) Pettitt, B. M.; Rossky, P. J. Alkali halides in water: ion-solvent correlations and ion-ion potentials of mean force at infinite dilution. *J. Chem. Phys.* **1986**, *84* (10), 5836–44.

(54) Akiyama, R.; Karino, Y.; Hagiwara, Y.; Kinoshita, M. Remarkable solvent effects on depletion interaction in crowding media: analyses using the integral equation theories. *J. Phys. Soc. Jpn.* **2006**, *75* (6), 064804/1–064804/7.

(55) Pratt, L. R.; Chandler, D. Hydrophobic solvation of nonspherical solutes. *J. Chem. Phys.* **1980**, *73* (7), 3430–3.

(56) Kovalenko, A.; Ten-No, S.; Hirata, F. Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *J. Comput. Chem.* **1999**, *20* (9), 928–936.

(57) Hamilton, T. P.; Pulay, P. Direct inversion in the iterative subspace (DIIS) optimization of open-shell, excited-state, and small multiconfiguration SCF wave functions. *J. Chem. Phys.* **1986**, *84* (10), 5728–34.

(58) Pulay, P. Improved SCF convergence acceleration. *J. Comput. Chem.* **1982**, *3* (4), 556–60.

(59) Pulay, P. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Lett.* **1980**, *73* (2), 393–8.

(60) Kelley, C. T. *Solving Nonlinear Equations with Newton's Method*; Society for Industrial and Applied Mathematics: Philadelphia, 2003; p 118.

(61) Perkyns, J.; Pettitt, B. M. A site-site theory for finite concentration saline solutions. *J. Chem. Phys.* **1992**, *97* (10), 7656–7666.

(62) Perkyns, J.; Pettitt, B. M. A dielectrically consistent interaction site theory for solvent-electrolyte mixtures. *Chem. Phys. Lett.* **1992**, *190* (6), 626–630.

(63) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic: London, 1976.

(64) Hirata, F.; Pettitt, B. M.; Rossky, P. J. Application of an extended RISM equation to dipolar and quadrupolar fluids. *J. Chem. Phys.* **1982**, *77* (1), 509–20.

(65) Pettitt, B. M.; Rossky, P. J. Integral equation predictions of liquid state structure for waterlike intermolecular potentials. *J. Chem. Phys.* **1982**, *77* (3), 1451–7.

(66) Hirata, F.; Rossky, P. J. An extended RISM equation for molecular polar fluids. *Chem. Phys. Lett.* **1981**, *83* (2), 329–34.

(67) Perkyns, J. S.; Pettitt, B. M. A dielectrically consistent interaction-site theory for solvent-electrolyte mixtures. *Chem. Phys. Lett.* **1992**, *190* (6), 626–30.

(68) Chandler, D.; Silbey, R.; Ladanyi, B. M. New and proper integral equations for site-site equilibrium correlations in molecular fluids. *Mol. Phys.* **1982**, *46* (6), 1335–45.

(69) Rossky, P. J.; Chiles, R. A. A complete integral equation formulation in the interaction site formalism. *Mol. Phys.* **1984**, *51* (3), 661–74.

(70) Gillan, M. J. A new method of solving the liquid structure integral equations. *Mol. Phys.* **1979**, *38* (6), 1781–1794.

(71) Saad, Y.; Schultz, M. H. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.* **1986**, *7* (3), 856–869.

(72) Booth, M. J.; Schlijper, A. G.; Scales, L. E.; Haymet, A. D. J. Efficient solution of liquid state integral equations using the Newton-GMRES algorithm. *Comput. Phys. Commun.* **1999**, *119* (2−3), 122–134.

(73) Kelley, C. T. A fast multilevel algorithm for integral equations. *SIAM J. Numer. Anal.* **1995**, *32* (2), 501–513.

(74) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–71.

(75) Dang, L. X.; Pettitt, B. M.; Rossky, P. J. On the correlation between like ion pairs in water. *J. Chem. Phys.* **1992**, *96* (5), 4046–7.

(76) Lee, C. Y.; McCammon, J. A.; Rossky, P. J. The structure of liquid water at an extended hydrophobic surface. *J. Chem. Phys.* **1984**, *80* (9), 4448–55.

(77) Chandler, D.; Weeks, J. D.; Andersen, H. C. van der Waals picture of liquids, solids, and phase transformations. *Science* **1983**, *220* (4599), 787–94.

(78) Hirata, F.; Rossky, P. J.; Pettitt, B. M. The interionic potential of mean force in a molecular polar solvent from an extended RISM equation. *J. Chem. Phys.* **1983**, *78* (6, Pt. 2), 4133–44.

CT8002817

JCTC Journal of Chemical Theory and Computation

# Protein Folding Pathways Revealed by Essential Dynamics Sampling

Daniele Narzi,[†,§] Isabella Daidone,*[,†,||] Andrea Amadei,[‡] and Alfredo Di Nola[†]

*Department of Chemistry, University of Rome 'La Sapienza', P.le Aldo Moro 5, 00185 Rome, Italy, and Dipartimento di Scienze e Tecnologie Chimiche, University of Rome 'Tor Vergata', via della Ricerca Scientifica 1, I-00133 Rome, Italy*

Received May 8, 2008

**Abstract:** The characterization of the protein folding process represents one of the major challenges in molecular biology. Here, a method to simulate the folding process of a protein to its native state is reported, the essential dynamics sampling (EDS) method, and is successfully applied to detecting the correct folding pathways of two small proteins, the all-$\beta$ SH3 domain of Src tyrosine kinase transforming protein (SH3) and the $\alpha/\beta$ B1 domain of streptococcal protein G (GB1). The main idea of the method is that a subset of the natural modes of fluctuation in the native state is key in directing the folding process. A biased molecular dynamics simulation is performed, in which the restrained degrees of freedom are chosen among those obtained by a principal component, or essential dynamics, analysis of the positional fluctuations of the C$\alpha$ atoms in the native state. Successful folding is obtained if the restraints are applied only to the eigenvectors with lowest eigenvalues, representing the most rigid quasi-constraint motions. If the essential eigenvectors, the ones accounting for most of the variance, are used, folding is not successful. These results clearly show that the eigenvectors with lowest eigenvalues contain the main mechanical information necessary to drive the folding process, while the essential eigenvectors represent the large concerted motions which can occur without folding/unfolding the protein.

## 1. Introduction

Understanding protein folding mechanisms represents one of the major aims of biophysics and molecular biology. Information on the sequence of conformational steps that lead to the native structure from denaturated polypeptide chains is fundamental to shed light on protein folding mechanisms, on the effects of different physicochemical conditions and mutations. Many experimental and theoretical approaches for the study of protein folding have been developed.[1−8] Computational methods represent a valid tool in order to obtain atomic details of such a process, and molecular dynamics (MD) simulations are among the most used ones.[9−12] A limitation encountered using MD simulations is due to the time scale accessible to this methodology that is not comparable with the time scale of most folding processes (ms-s). At present, with standard MD simulations this process can be well simulated only for short peptides[9,11,13,14] but is still beyond reach for globular proteins. To overcome this problem different MD techniques have been developed.

The simplest approach is to perform high-temperature MD simulations starting from the native structure to study the unfolding process.[15−22] In some instances, by considering the unfolding as the reverse of folding, information on the folding process is inferred from the high-temperature unfolding simulations.[20−22] A more sophisticated approach makes

* Corresponding author e-mail: Isabella.Daidone@iwr.uni-heidelberg.de.

† University of Rome 'La Sapienza'.

§ Current address: Theoretical & Computational Membrane Biology, Center for Bioinformatics Saar, Universität des Saarlandes, D-66041 Saarbrücken, Germany.

|| Current address: Interdisciplinary Center for Scientific Computing, University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany.

‡ University of Rome 'Tor Vergata'.

10.1021/ct800157v CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/21/2008

use of unfolding simulations followed by the calculation of the free energy of the folding process at $T = 300$ K, by means of the umbrella sampling method, along the previously determined path. Unfolding is performed by high temperature[23,24] or by applying a harmonic potential to different reaction coordinates, such as the radius of gyration[25] and the end-to-end distance.[26,27] The previously reported approaches are based on the hypothesis that the folding process at 300 K follows the same path of the unfolding process performed with high temperature or with a harmonic potential, but the issue of whether unfolding simulations are representative for the folding process is still open.[28,29]

Different from the previous methods, in 'targeted molecular dynamics' (TMD) simulation,[30] a folding simulation is performed along a path not previously determined. This is accomplished by applying a harmonic, time-dependent, restraint on each atom to continuously decrease the all-atom root-mean-square deviation from the native state. Other methods make use of simplified molecular models in order to gain computation time by neglecting details. This category includes the widely used so-called lattice models.[31,32] Additionally, accurate prediction of native three-dimensional protein structures could be reached using semiempirical database-driven prediction methods.[33−35]

Here, the essential dynamics sampling (EDS) method[36,37] is used to simulate protein folding. Starting from an unfolded structure, a usual MD simulation step is performed. The new structure is accepted only if its distance from the native structure does not increase, otherwise, it is projected onto the closest configuration having the same distance to the native conformation as the structure before the MD step. The distance is calculated in a configurational subspace defined by a set of generalized coordinates obtained by a principal component, or essential dynamics, analysis[38−40] of a native-state equilibrium simulation. Hence, correct folding can be obtained by using only a small fraction of the degrees of freedom of the protein to bias the MD simulation toward its native conformation.

Due to the absence of any restraining potential, as used for example in TMD, the EDS method does not force the system to overcome barriers higher than a few kT. Thus, the protein is not allowed to undergo major unfolding if incorrect packing leads to kinetic traps or, in other words, to off-pathway intermediates. In this sense the method is somewhat similar to the CONTRA MD[41] but differs mainly in the choice of the reaction coordinates which, in the present case, are chosen so to contain information on the dynamical properties of the native state and might, hence, represent better candidates than reaction coordinates often used such as the radius of gyration or the root-mean-square deviation from the native structure.

The EDS method was successfully applied to the folding process of cytochrome *c*,[42] an all-α protein. In the present work, the method is further extended, and its ability in reproducing the native conformation and the known folding steps in proteins with different topologies, namely the all-$\beta$ and $\alpha/\beta$ motives, is verified. The model systems used are the SH3 domain of Src tyrosine kinase transforming protein (SH3) and the B1 domain of streptococcal protein G (GB1). The src-SH3, a 56-residue all-$\beta$ protein, was largely investigated by MD simulations,[43−46] and, in agreement with experimental data,[47,48] the folding transition state is found to be characterized by the presence of the central three-stranded $\beta$-sheet, whereas the formation of the hydrophobic sheet, consisting of the two terminal strands, is observed in the last stage of the folding process. The GB1, a 56 residues $\alpha/\beta$ protein, has been shown to populate an intermediate state along its folding process with native-like structural elements involving one of the four strands, namely the $\beta$3 strand.[49,50]

The results of the EDS folding simulations performed here show that in SH3 the central three-stranded $\beta$-sheet precedes the whole structure formation and in GB1 the native-contacts formation of the $\beta$3 strand is a prerequisite for a correct folding. These results are in agreement with experiments,[47−50] thus assessing the predictive capabilities of the method.

## 2. Methods

**2.1. Molecular Dynamics Simulations.** All MD simulations were performed using the GROMACS software package and the Gromos87 force field[51] with modification as suggested by van Buuren et al.[52] In both cases the proteins were solvated with water in a periodic cubic box of dimensions $57.0 \times 57.0 \times 57.0$ Å. The simple point charge[53] water model was used. Neutralization of the total charge of the system was obtained by replacing 3 and 4 molecules of water with 3 and 4 Na ions for SH3 and GB1, respectively. The SHAKE algorithm[54] was used to constrain all bond lengths. A time step of 2 fs was used for numerical integration. The isokinetic temperature coupling[55] was used to keep the temperature constant. The long-range electrostatic interactions were treated with the particle mesh Ewald method[56] using a $48 \times 48 \times 48$ grid combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points, whereas the short-range electrostatic interactions were treated with a nonbonded pair-list cutoff of 9.0 Å.

**2.2. Native-State Simulations and Essential Dynamics (ED) Analysis.** For both SH3 and GB1 a 5000 ps long MD simulation of the native state was performed at room temperature ($T = 300$ K) in the NVT ensemble (at a liquid density of 55.32 mol/L). The starting structures were taken from the NMR structure (PDB entry 1srl)[57] for SH3 and from the 2.07 Å resolution refined crystal structure (PDB entry 1pga)[58] for GB1. From the equilibrated portion of the native-state trajectory (beyond 200 ps) the covariance matrix of the positional fluctuations of the Cα carbon atoms was built up and diagonalized. The procedure yields new axes (eigenvectors), representing the directions of the concerted motions. The corresponding eigenvalues give the mean square positional fluctuation for each direction.[38,39] 168 eigenvectors were obtained for each protein, corresponding to the number of degrees of freedom of the Cα carbon atoms. Sorting the eigenvectors by the size of the eigenvalues shows that the configurational space can be divided in a low dimensional (essential) subspace (the first 10−15 eigenvectors in the present proteins) in which most of the positional fluctuations are confined ($\approx$ 60−70% of the total variance) and a high

dimensional (near-constraint) subspace in which small-amplitude fluctuations occur.

**2.3. Essential Dynamics Sampling (EDS).** The essential dynamics sampling technique[36,37,59] can be used to decrease the distance of a given structure from a reference structure in a space defined by a subset of eigenvectors as obtained by the ED analysis of the native-state MD simulation (see the Results section for the choice of the set of eigenvectors used in the present work).

In the EDS simulation a usual MD simulation step is performed starting from an unfolded conformation; at each step the distance from the reference conformation (the crystal or NMR structure in the present cases) is calculated in the chosen subspace. If this distance does not increase, the new conformation is accepted. Otherwise, the coordinates (in the chosen subspace) are radially corrected in order to keep the position onto the hypersphere centered on the reference conformation, with a radius given by the distance from the reference in the previous step. This correction step is performed using a nonstationary holonomic constraint in the chosen subspace $\xi$

$$G(\xi(t+\Delta t);t+\Delta t) = |\xi(t+\Delta t) + \Delta\xi_c - \xi_0|^2 - |\xi(t) - \xi_0|^2 = 0 \quad (1)$$

where $\xi(t)$ and $\xi(t+\Delta t)$ are the unconstrained positions at time $t$ and $(t+\Delta t)$, respectively, $\Delta\xi_c$ is the correction for the application of the constraint, and $\xi_0$ is the reference position (the crystal or NMR structure). Eq 1 does not suffice to solve for $\Delta\xi_c$ in a unique way. To obtain a unique solution, we add the requirement that $|\Delta\xi_c|^2$ is minimized. This is achieved using one Lagrangian multiplier:

$$\Delta\xi_c^i - \lambda\frac{\partial G}{\partial\Delta\xi_c^i} = \Delta\xi_c^i - 2\lambda(\xi^i(t+\Delta t) + \Delta\xi_c^i - \xi_0^i) = 0 \quad (2)$$

and therefore

$$\Delta\xi_c^i = \frac{2\lambda}{1-2\lambda}(\xi^i(t+\Delta t) - \xi_0^i) \quad (3)$$

Using eq 3 and eq 1 $\lambda$ can be expressed as a function of $\xi(t)$, $\xi(t+\Delta t)$, and $\xi_0$. This value of $\lambda$, and the corresponding $\Delta\xi_c$, is then used to correct $\xi(t+\Delta t)$ to fulfill the constraint with the least perturbation.

**2.4. Unfolding/Folding Simulations.** Starting from two different structures extracted from the native-state MD simulations at $t$ = 2000 ps and $t$ = 3000 ps, two high temperature unfolding simulations of 3500−4000 ps were performed for each protein. The temperature was kept at a value of 500 K, and the system was coupled to a pressure bath at a value of 1 bar. It has to be pointed out that these conditions are not meant to represent a real unfolding process, and the corresponding trajectories are not used for analysis purpose. They are rather meant as a computational procedure to generate a large number of denatured structures to be used as starting points in the folding simulations. For the GB1 protein a further simulation of 5000 ps was performed coupling the residues corresponding to the $\beta$3 strand (GLY41-ASP47) at a temperature of 300 K and the remaining residues at 500 K (see the Results section for the justification of this simulation).

Six and twelve protein structures for the SH3 and the GB1, respectively, were extracted from these unfolding simulations and used as starting structures in the EDS folding simulations. The selected structures are characterized by high values of root-mean-square deviation with respect to either the NMR or the crystal structure and high radius of gyration. Nevertheless, they retain some degree of secondary structure (see the Results section). Experimental and computational methods have demonstrated that even under strong denaturing conditions unfolded structures retain a residual native-like secondary structure.[60,61] Therefore, we believe that the unfolded structures used in the present work, which in fact contain information from the native starting conformation, are good candidates as representative structures of the unfolded ensemble.

The starting unfolded structures were solvated in water and equilibrated for 10 ps at a temperature of 300 K and a pressure of 1 bar. The folding simulations were then performed in the NVT ensemble at room temperature ($T$ = 300 K). A slightly different procedure is used in the present work, with respect to the one previously reported:[42] to allow a local increase of the distance from the reference, each 10 ps of EDS simulation is followed by 10 ps of unbiased MD simulation.

## 3. Results

**3.1. EDS Procedure.** Preliminary analyses were performed to assess the relevance of using different sets of the native eigenvectors, accounting for the Cα carbon atoms fluctuations, in the biasing procedure of the folding simulations (only results for the GB1 are reported here since similar results are also obtained for SH3).

Starting from an unfolded conformation (structure RUN1 in Figure 1) three initial folding simulations were performed using all the Cα eigenvectors, the high-variance essential eigenvectors (the first 13), and the low-variance eigenvectors (the last 155) - $RUN1_{all}$, RUN1′, and RUN1, respectively, in Table 1. At the end of both simulations that included the essential eigenvectors a compact structure is reached, but almost no secondary and tertiary structure is recovered (see $RUN1_{all}$ and RUN1′ in Table 1).

In order to characterize the two sets of the Cα eigenvectors, i.e., providing and not providing correct folding, the nature of the associated motions was investigated. For this purpose, the overall displacement of the Cα atoms belonging to a given secondary-structure element was decomposed into internal motions, i.e., occurring within the secondary structure, and roto-translational motion, i.e., of the secondary-structure element with respect to its Cα centroid. An example for the $\beta$3-$\beta$4 sheet of GB1 is reported in Figure 2. The results make evidence that the last, i.e., with the lowest eigenvalues, 150−155 eigenvectors (out of 168) mostly represent internal collective vibrations, i.e., within the $\beta$-sheet, whereas the essential eigenvectors (the first 10−20) mainly provide roto-translational motions of the $\beta$-sheet.

These results show that the quasi-constraint, low-variance eigenvectors, that were shown here to represent in the folded protein the smallest vibrations within each secondary struc-

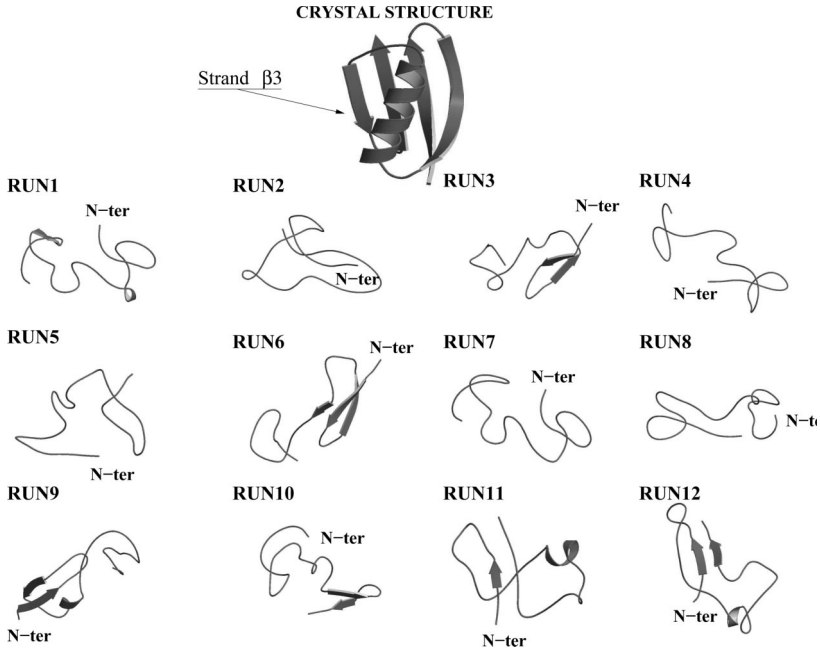



***Figure 1.*** GB1. Backbone conformation of the crystal structure and of the twelve structures used as starting structures in the EDS folding simulations.

***Table 1.*** GB1: Starting and Final Radius of Gyration ($Rg_i$, $Rg_f$), Backbone Root-Mean-Square Deviation ($RMSD_i$, $RMSD_f$) and Native Contact Fraction ($\rho_i$, $\rho_f$) with Respect to the Crystal Structure, Number of Residues in $\beta$-Structure ($N_{\beta_i}$, $N_{\beta_f}$) and in $\alpha$-Structure ($N_{\alpha_i}$, $N_{\alpha_f}$) in the Folding Simulations[a]

| SIM | $Rg_i$ (nm) | $Rg_f$ (nm) | $RMSD_i$ (nm) | $RMSD_f$ (nm) | $\rho_i$ | $\rho_f$ | $N_{\beta_i}$ | $N_{\beta_f}$ | $N_{\alpha_i}$ | $N_{\beta_f}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| crystal | 1.05 | | – | | – | | 24 | | 14 | |
| NatGB1 | 1.05(0.01) | | 0.12(0.02) | | 0.91(0.02) | | 23(1) | | 14(1) | |
| RUN1$_{all}$ | 1.28 | 1.07 | 1.14 | 0.15 | 0.25 | 0.51 | 0 | 10 | 0 | 7 |
| RUN1′ | 1.28 | 1.06 | 1.14 | 0.15 | 0.25 | 0.48 | 0 | 8 | 0 | 5 |
| RUN1 | 1.28 | 1.07 | 1.14 | 0.13 | 0.25 | 0.89 | 0 | 21 | 0 | 15 |
| RUN2 | 1.31 | 1.05 | 1.19 | 0.10 | 0.18 | 0.92 | 8 | 22 | 4 | 15 |
| RUN3 | 1.27 | 1.05 | 0.97 | 0.14 | 0.40 | 0.84 | 6 | 23 | 9 | 14 |
| RUN4 | 1.47 | 1.05 | 1.21 | 0.12 | 0.28 | 0.88 | 3 | 16 | 2 | 15 |
| RUN5 | 1.30 | 1.09 | 1.15 | 0.32 | 0.25 | 0.61 | 0 | 7 | 5 | 0 |
| RUN6 | 1.37 | 1.03 | 1.03 | 0.15 | 0.32 | 0.86 | 15 | 16 | 0 | 14 |
| RUN7 | 1.27 | 1.04 | 1.12 | 0.13 | 0.27 | 0.79 | 2 | 19 | 6 | 11 |
| RUN8 | 1.32 | 1.04 | 1.16 | 0.20 | 0.18 | 0.75 | 0 | 14 | 2 | 14 |
| RUN9 | 1.31 | 1.02 | 1.02 | 0.17 | 0.37 | 0.85 | 6 | 12 | 4 | 15 |
| RUN10 | 1.19 | 1.06 | 1.24 | 0.27 | 0.27 | 0.71 | 15 | 4 | 0 | 15 |
| RUN11 | 1.24 | 1.05 | 1.10 | 0.15 | 0.44 | 0.85 | 7 | 20 | 0 | 13 |
| RUN12 | 1.28 | 1.05 | 1.13 | 0.08 | 0.35 | 0.90 | 10 | 24 | 0 | 15 |
| RUN4′ | 1.47 | 1.07 | 1.21 | 0.14 | 0.28 | 0.82 | 3 | 21 | 2 | 15 |
| RUN8′ | 1.32 | 1.06 | 1.16 | 0.13 | 0.18 | 0.81 | 0 | 20 | 2 | 14 |

[a] The final values in the folding simulations are averaged over the last 100 ps of each simulation. The values for the native-state trajectory (NatGB1) are averaged on the equilibrated part (200−5000 ps) with standard deviations in parentheses. The number of Cα eigenvectors used in the EDS procedure is as follows: all in RUN1$_{all}$; the first 13 in RUN1′; the last 155, i.e., the last 90%, in RUN1-RUN12; the last 90% of the eigenvectors calculated including not only all the Cα atoms but also the side-chain atoms of residues 41−47 in RUN4′ and RUN8′.

ture element, contain the proper mechanical information for the folding process, whereas the essential eigenvectors represent the large collective motions which can occur without folding/unfolding the protein.

It should be noted that in the previous study on the cytochrome *c*,[42] an all-α protein, a correct folding of the protein was obtained performing EDS folding simulations on a smaller space with respect to the GB1 and SH3, i.e., the last 30% of the eigenvectors versus the last 90% used here. When only the last 30% of the eigenvecors was used for GB1 and SH3, folding was not successful (data not shown). We assign this difference to the fact that $\beta$ or $\alpha/\beta$ folds, such as SH3 and GB1, are characterized by higher contact order with respect to α topologies, such as cytochrome *c*, and hence the main mechanical information necessary for folding is distributed over a larger number of degrees of freedom.

In what follows we will perform different independent folding simulations using the last 155 eigenvectors for the GB1 and the last 160 for SH3.

**3.2. GB1.** The main structural properties of the native-state MD simulation at 300 K (NatGB1) are reported in Table 1. The data show a good agreement with the crystal structure. The values of the radius of gyration (Rg), root-mean-square

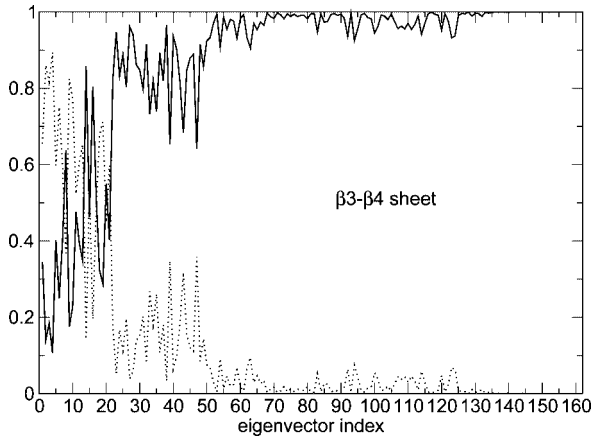



***Figure 2.*** Fraction of internal (solid line) and roto-translational, with respect to the Cα atoms centroid (dotted line), displacements of the Cα atoms due to the motion along each eigenvector for the $\beta$3-$\beta$4 sheet of GB1. The following procedure was used: configurations of the secondary structure element of interest, as obtained by the filtered motion of a given eigenvector, were least-squares fitted to the corresponding average configuration. The mean square fluctuation recalculated for the given secondary-structure element after this procedure provides the internal-motion contribution to the total mean square fluctuation of the secondary structure element due to the eigenvector motions, while the residual fluctuation is ascribed to the roto-translational motion of the secondary structure element.

deviation (RMSD), % of native contacts ($\rho$) with respect to the crystal structure, and number of residues in $\beta$- and α-structure ($N_\beta$ and $N_\alpha$) of the ten unfolded structures used as starting points for the folding simulations are reported in Table 1 as well. All the starting structures are characterized by high RMSD and Rg and low native contacts and secondary structure contents. The corresponding conformations are shown in Figure 1, together with the crystal structure.

To simulate the folding process, the EDS was performed for a time range of 3000−5000 ps for each starting structure in a subspace defined by the last 155 eigenvectors of the covariance matrix of the Cα positional fluctuations (see the 'EDS Procedure' section). The final structural properties, averaged over the last 100 ps of each folding simulation, are reported in Table 1. Although from the table it is not completely clear which simulations are really successful, further analyses (vide infra) suggest that three simulations (RUN1−3) out of ten were successful.

The side-chain RMSD, with respect to the crystal structure, averaged on the three EDS simulations providing the correctly folded structures and on the last 100 ps of each simulation, is reported in Figure 3. A good agreement with the RMSD calculated on the native-state trajectory, reported in the same figure, can be observed. This result shows that, although the constraint applied in EDS accounts only for Cα atoms, the correct conformation of the side chains was obtained in EDS folding simulations.

The analysis of the trajectories shows that to achieve the final correct folding of the $\beta$3-$\beta$4 sheet, the TRP43 and TYR45 side chains (belonging to the $\beta$3 strand shown in

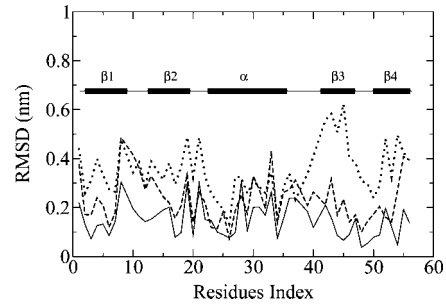



***Figure 3.*** GB1. Side-chain RMSD with respect to the crystal structure. Solid line: average on the equilibrated part of the native-state simulation (200−5000 ps). Dashed line: average over the last 100 ps of the three correctly folded simulations (RUN1−3). Dotted line: average over the last 100 ps of the seven not correctly folded simulations (RUN4−10).

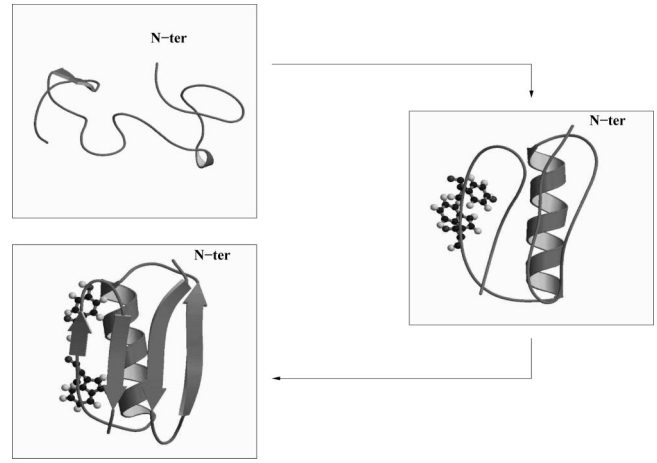



***Figure 4.*** GB1. Backbone conformations at $t = 0$ ps, 984 ps, 4000 ps for RUN1. The side-chain orientations of TRP43 and TYR45 are also reported at $t = 984$ ps and $t = 4000$ ps.

Figure 1) need to be oriented toward the α-helix. As an example in Figure 4 we report representative structures along RUN1, the initial and final conformations, and a conformation observed at an intermediate time ($t = 984$ ps), in which the TRP43 and TYR45 side chains, highlighted in the figure, both point at the interface with the α-helix.

Conversely, at the end of the seven nonsuccessful simulations the region corresponding to the $\beta$3 strand (residues 41−47) is characterized by very high values of the side-chain RMSD with respect to the crystal structure (Figure 3). In particular all seven final structures show that the TRP43 and/or TYR45 side chains point away from the α-helix, preventing the correct folding of the $\beta$3-$\beta$4 sheet, as shown in Figure 5.

These results are in agreement with experimental data,[62−64] indicating that the native-state fluorescence intensity of TRP43 is recovered more rapidly than the formation of stable hydrogen bonds in the $\beta$-sheets, thus implying that rapid partial or complete formation of the tertiary contacts between the $\beta$3-$\beta$4 sheet and the α-helix occurs.

To further verify this hypothesis two different strategies were used. In the first one the purpose was to obtain additional unfolded conformations with some native-like structural properties for the residues corresponding to the

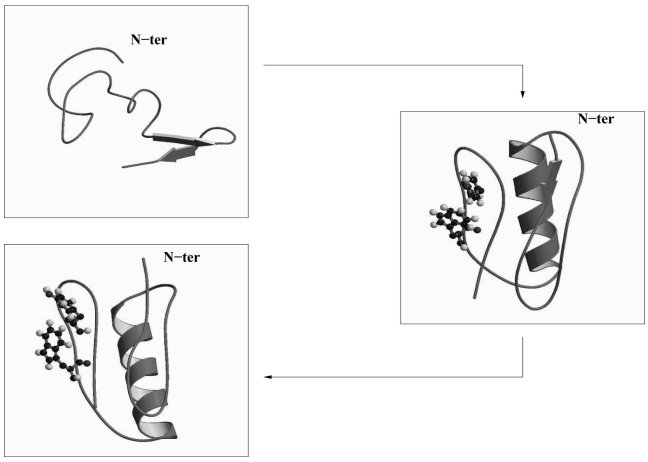



***Figure 5.*** GB1. Backbone conformations at $t = 0$ ps, 746 ps, 4000 ps for RUN10. The side-chain orientations of TRP43 and TYR45 are also reported at $t = 746$ ps and $t = 4000$ ps.

$\beta$3 strand (residues 41−47). To this end a further unfolding simulation was performed starting from the structure extracted at $t = 4000$ ps of the native-state simulation and coupling the residues of the $\beta$3 strand to a thermal bath at $T = 300$ K, while the rest of the system was kept at $t = 500$ K. Two unfolded structures were extracted from this simulation and used as starting structures in the folding process (RUN11−12). Their conformations are shown in Figure 1. The structural properties of the starting and final conformations are reported in Table 1. A good agreement with the values obtained in the native-state simulation can be observed, thus indicating that the correct fold is obtained.

The second strategy consisted of using a new set of eigenvectors for the EDS simulations which included the Cα atoms of the whole protein together with the side-chain atoms of residues 41−47. Then, new simulations starting from the initial structures of the unsuccessful RUN4 and RUN8 using these new eigenvectors were performed. The results reported in Table 1 (RUN4′ and RUN8′) show that a correct folding was obtained in both cases.

**3.3. SH3.** The main structural properties of the native-state MD simulation at 300 K (NatSH3) are reported in Table 2. The data show a good agreement with the NMR structure. In the table are also reported the RMSD with respect to the NMR structure, Rg, $N_\beta$, and $\rho$ values of the six unfolded structures used as starting points for the folding simulations. Figure 6 shows the corresponding structures together with the NMR one.

For each starting structure, the folding process was simulated for a time range of 3000−5000 ps by the EDS performed in a subspace defined by the last 160 eigenvectors over a total of 168 obtained from the native-state simulation (see the 'EDS Procedure' section). The final structural properties, averaged on the last 100 ps of each folding simulation, are reported in Table 2. The correct conformation was reached in five simulations (RUN1−5), that can be considered representative of the folding process, whereas the structure obtained in the sixth one showed values of RMSD, $\rho$, and $N_\beta$ not in agreement with the values obtained in the native simulation.

The side-chain RMSD with respect to the NMR structure, averaged on the last 100 ps of the five EDS trajectories providing the correctly folded structures, is reported in Figure 7 and compared with the RMSD in the native-state simulation averaged on the equilibrated part of the simulation (200−5000 ps). It results in a good agreement between the two curves with the exception of the side chain of THR42 (a residue forming a $\beta$-turn), that shows a larger RMSD value at the end of the folding simulations than in the native one.

The analysis of the native contacts as a function of time shows that in four simulations, out of five correctly folded, the native interactions within the central $\beta$-sheet, consisting of the $\beta$2, $\beta$3, and $\beta$4 strands, precede the ones of the terminal $\beta$-sheet, consisting of the $\beta$1 and $\beta$5 strands, in agreement with experimental[47,48] and computational[43,46] data. As an example, the native contact maps, calculated at different times along the RUN3 trajectory, are reported in Figure 8 and compared with the contact map obtained from the native-state trajectory. The maps were calculated averaging over 10 ps starting at time $t = 0$, 100, 200, 2800 ps. A native contact between non-neighboring residues was considered to be formed if at least one distance between any two atoms was smaller than 0.6 nm. In the starting structure ($t = 0$ ps in Figure 8) the $\beta$2-$\beta$3 interaction is partially present; however, it is not complete, and the secondary structure is only partially formed (see structure RUN3 in Figure 6). At $t = 100$ ps the $\beta$2-$\beta$3 interactions are completely formed as well as the secondary structure, and part of the $\beta$1-$\beta$5 and the $\beta$3-$\beta$4 contacts are present. At $t = 200$ ps the $\beta$1-$\beta$5 and the $\beta$3-$\beta$4 interactions are almost completely formed, and at $t = 2800$ ps the contact map is similar to the one calculated on the native-state trajectory.

## 4. Conclusions

In the present work the EDS method is used to fold two small proteins, the all-$\beta$ SH3 and the $\alpha/\beta$ GB1 protein, to their native structures starting from unfolded conformations and to reveal the known folding steps. The idea of the method is to bias the system toward its known native structure by means of a MD simulation, using a least biased procedure. This is accomplished by restraining only a subset of the degrees of freedom of the protein and by choosing such coordinates so to contain dynamical information of the native state. This is achieved by using a subgroup of the eigenvectors extracted from a principal component (or essential dynamics) analysis of the collective motions of the backbone Cα atoms of the protein in its native state. Hence, no information of the side chains is introduced. It is shown here that the EDS method does not "force", e.g., overcoming barriers higher than a few kT's, the simulation toward the correct folded structure; in fact not all the folding simulations were successful, in particular for the GB1 protein. When the protein gets into a nonproductive folding trap, the folded structure is not reached. Moreover, since the reaction coordinates used here contain information on the native state, it is possible with this procedure to find out the main mechanical information necessary for the folding process.

The results showed that in SH3 the native interactions within the central $\beta$-sheet precede the ones of the terminal

***Table 2.*** SH3: Starting and Final Radius of Gyration ($Rg_i$, $Rg_f$), Backbone Root-Mean-Square Deviation ($RMSD_i$, $RMSD_f$) and Native Contact Fraction ($\rho_i$, $\rho_f$) with Respect to the Crystal Structure, Number of Residues in $\beta$-Structure ($N_{\beta i}$, $N_{\beta f}$) in the Folding Simulations[a]

| SIM | $R_{gi}$ (nm) | $R_{gf}$ (nm) | $RMSD_i$ (nm) | $RMSD_f$ (nm) | $\rho_i$ | $\rho_f$ | $N_{\beta_i}$ | $N_{\beta_f}$ |
|---|---|---|---|---|---|---|---|---|
| NMR | 1.04 | | – | | – | | 21 | |
| NatSH3 | 1.01(0.01) | | 0.19(0.03) | | 0.89(0.02) | | 22(2) | |
| RUN1 | 1.49 | 1.03 | 1.24 | 0.11 | 0.35 | 0.88 | 10 | 19 |
| RUN2 | 1.29 | 1.03 | 1.05 | 0.19 | 0.27 | 0.86 | 11 | 20 |
| RUN3 | 1.29 | 1.02 | 0.94 | 0.17 | 0.30 | 0.88 | 11 | 21 |
| RUN4 | 1.33 | 1.03 | 1.22 | 0.10 | 0.21 | 0.85 | 0 | 20 |
| RUN5 | 1.35 | 1.02 | 0.99 | 0.13 | 0.36 | 0.83 | 10 | 20 |
| RUN6 | 1.34 | 1.08 | 1.35 | 0.47 | 0.16 | 0.50 | 5 | 7 |

[a] The final values in the folding simulations are averaged over the last 100 ps of each simulation. The values for the native-state trajectory (NatSH3) are averaged on the equilibrated part (200−5000 ps) with standard deviations in parentheses. In all folding simulations (RUN1-RUN6) the last 160 Cα eigenvectors are used in the EDS procedure.

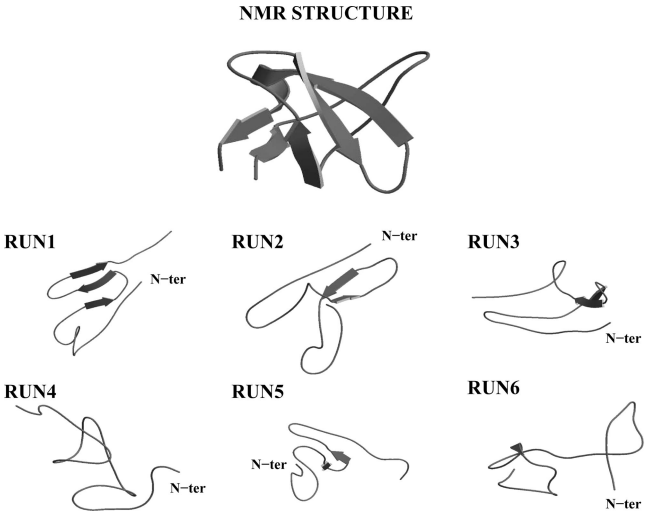



***Figure 6.*** SH3. Backbone conformation of the NMR structure and of the six structures used as starting structures in the EDS folding simulations.

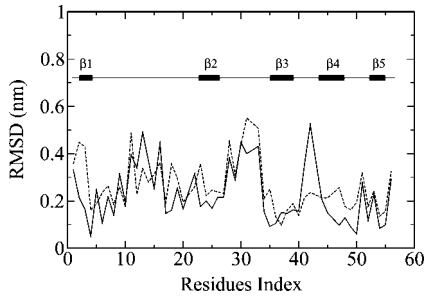



***Figure 7.*** SH3. Side-chain RMSD with respect to the NMR structure. Solid line: average over the last 100 ps of the five correctly folded simulations (RUN1−5). Dashed line: average on the equilibrated part of the native-state simulation (200−5000 ps).

$\beta$-sheet, in agreement with experimental[47,48] and computational[43,46] data. In GB1, a correct folding of the side chains of TRP43 and TYR45 is a prerequisite for a correct folding, in agreement with experimental data[62−64] that show that the native-state fluorescence intensity of TRP43 is recovered more rapidly than the formation of stable $\beta$-sheet hydrogen bonds. These results, together with those previously reported for cytochrome *c*, confirm that EDS can detect the main structural characteristics of the folding mechanism. In this sense it could be used to predict crucial interactions in the folding of proteins, although validation from experiments is required.

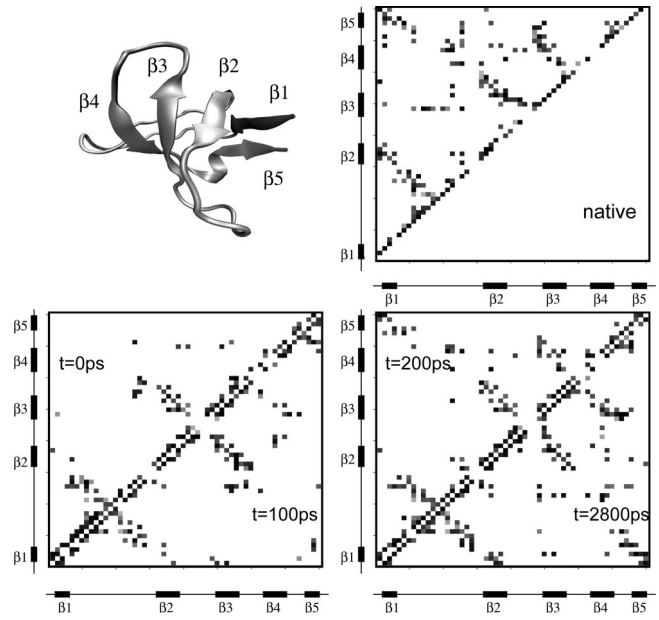



***Figure 8.*** SH3. Native-contact maps. (top left side) Backbone conformation of the NMR structure showing the five $\beta$-strands: $\beta$1 residues 2−4, $\beta$2 residues 23−26, $\beta$3 residues 35−39, $\beta$4 residues 44−48, $\beta$5 residues 53−55. (top right side) Native-contact map obtained from the native-state trajectory. (Bottom) Native-contact maps at different times along RUN3, each calculated by averaging over 10 ps starting at $t = 0$ ps, 100 ps, 200 ps, and 2800 ps.

**Acknowledgment.** This work was supported by the Italian FIRB RBIN04PWNC_001 "Structure, function, dynamics and folding of proteins" founded by MIUR. We also acknowledge the University of Rome 'La Sapienza' for financial support with the project "MORFOGENESI MOLECOLARE: un approccio multidisciplinare per lo studio del folding e misfolding delle proteine" and CASPUR (Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Universitá e Ricerca) for the use of its computational facilities.

## References

(1) McCammon, J. A.; Gelin, B.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(2) Wong, C. F.; Zheng, C.; Shen, J.; McCammon, J. A.; Wolynes, P. G. *J. Phys. Chem.* **1993**, *97*, 3100–3110.

(3) Dill, K.; Chan, H. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

(4) Dobson, C. M.; Karplus, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92–101.

(5) Onuchic, J. N.; Nymeyer, H.; García, A. E.; Chahine, J.; Socci, N. D. *Adv. Protein Chem.* **2000**, *53*, 87–152.

(6) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191–2195.

(7) Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1849–1857.

(8) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.

(9) Daura, X.; Gademann, K.; Juan, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236–240.

(10) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.

(11) Daidone, I.; D'Abramo, M.; Di Nola, A.; Amadei, A. *J. Am. Chem. Soc.* **2005**, *127*, 14825–14832.

(12) Perez, A.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(13) Daidone, I.; Amadei, A.; Di Nola, A. *Proteins* **2005**, *59*, 510–518.

(14) Daidone, I.; Ulmschneider, M. B.; Di Nola, A.; Amadei, A.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15230–15235.

(15) Tirado-Rives, J.; Jorgensen, W. L. *Biochemistry* **1991**, *30*, 3864–3861.

(16) Tirado-Rives, J.; Jorgensen, W. L. *Biochemistry* **1993**, *32*, 4175–4184.

(17) Caflish, A.; Karplus, M. *J. Mol. Biol.* **1995**, *252*, 672–708.

(18) Lazaridis, T.; Lee, I.; Karplus, M. *Protein Sci.* **1997**, *6*, 2589–2605.

(19) Li, A. J.; Daggett, V. *J. Mol. Biol.* **1998**, *275*, 677–694.

(20) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518–13522.

(21) Alonso, D. O. V.; Daggett, V. *Proc. Natl. Acad. Sci U.S.A.* **2000**, *97*, 133–138.

(22) Pan, Y.; Daggett, V. *Biochemistry* **2001**, *40*, 2723–2731.

(23) Sheinerman, F. B.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1562–1567.

(24) Sheinermann, F. B.; Brooks, C. L., III *J. Mol. Biol.* **1998**, *278*, 439–456.

(25) Marchi, M.; Ballone, P. *J. Chem. Phys.* **1999**, *110*, 3697–3702.

(26) Paci, E.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 441–459.

(27) Paci, E.; Smith, L. J.; Dobson, C. M.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 329–347.

(28) Finkelstein, A. V. *Protein Eng.* **1997**, *10*, 843–845.

(29) Wang, T.; Wade, R. C. *J. Chem. Theory Comput.* **2007**, *3*, 1476–1483.

(30) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2000**, *39*, 252–260.

(31) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. *Fold. Des.* **1998**, *3*, 183–194.

(32) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544–2549.

(33) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209–225.

(34) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B.; Bystroff, C.; Baker, D. *Proteins* **1999**, *34*, 82–95.

(35) Bonneau, R.; Strauss, C. E.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmstrom, L.; Robertson, T.; Baker, D. *J. Mol. Biol.* **2002**, *322*, 65–78.

(36) Amadei, A.; Linssen, A. B. M.; de Groot, B. L.; van Aalten, D. M.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13*, 615–625.

(37) de Groot, B. L.; Amadei, A.; van Aalten, D. M. F.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13*, 741–751.

(38) García, A. E. *Phys. Rev. Lett.* **1992**, *66*, 2696–2699.

(39) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.

(40) Meyer, T.; Ferrer-Costa, C.; Perez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. *J. Chem. Theory Comput.* **2006**, *2*, 251–258.

(41) Harvey, S. C.; Gabb, H. A. *Biopolymers* **1993**, *13*, 741–751.

(42) Daidone, I.; Amadei, A.; Roccatano, D.; Di Nola, A. *Biophys. J.* **2003**, *85*, 2865–2871.

(43) Shea, J. E.; Onuchic, J. N.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16064–16068.

(44) Guo, W.; Lampoudi, S.; Shea, J. E. *Biophys. J.* **2003**, *85*, 61–69.

(45) Guo, W.; Lampoudi, S.; Shea, J. E. *Proteins* **2004**, *55*, 395–406.

(46) Ding, F.; Guo, W.; Dokholyan, N. V.; Shakhnovich, E. I.; Shea, J. E. *J. Mol. Biol.* **2005**, *350*, 1035–1050.

(47) Riddle, D. S.; Grantcharova, P. V.; Santiago, J. V.; Alm, E.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016–1024.

(48) Grantcharova, P. V.; Riddle, D. S.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *13*, 7084–7089.

(49) McCallister, L. E.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669–673.

(50) Nauli, S.; Kuhlman, B.; Baker, D. *Nat. Struct. Biol.* **2001**, *8*, 602–605.

(51) van Gunsteren, W. F.; Berendsen, H. J. C. *Gromos manual*; BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen: The Netherlands, 1987.

(52) van Buuren, A. R.; Marrink, S. J.; Berendsen, H. J. C. *J. Phys. Chem.* **1993**, *97*, 9206–9212.

(53) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981.

(54) Ryckaert, J. P.; Bellemans, A. *Chem. Phys. Lett.* **1975**, *30*, 123–125.

(55) Brown, D.; Clarke, J. H. R. *Mol. Phys.* **1984**, *51*, 1243–1252.

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

(57) Yu, H.; Rosen, M. K.; Schreiber, S. L. *FEBS Lett.* **1993**, *324*, 87–92.

(58) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4720.

(59) Roccatano, D.; Daidone, I.; Ceruso, M.-A.; Bossa, C.; Di Nola, A. *Biophys. J.* **2003**, *84*, 1876–1883.

(60) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.

(61) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153–164.

(62) Park, S.; O'Neil, K. T.; Roder, H. *Biochemistry* **1997**, *36*, 14277–14283.

(63) Park, S.; Ramachandra Shastry, M. C.; Roder, H. *Nat. Struct. Biol.* **1999**, *6*, 943–947.

(64) Kuszewski, J.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1994**, *3*, 1945–1952.

CT800157V

JCTC Journal of Chemical Theory and Computation

# Direct Comparison of Experimental and Calculated NMR Scalar Coupling Constants for Force Field Validation and Adaptation

Franziska F.-F. Schmid and Markus Meuwly*

*Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland*

Received June 24, 2008

**Abstract:** The ability to measure scalar coupling constants across hydrogen bonds ($^{3h}J_{NC'}$) from high-resolution NMR experiments allows the characterization of detailed structural properties of biomolecules. To analyze those, a parametrized model based on the linear combination of atomic orbitals relates H-bond geometries with the measured $^{3h}J_{NC'}$ coupling magnitude. In the present study the dependence of calculated $^{3h}J_{NC'}$ coupling constants on force field parameters is assessed. It is shown that increased polarity of the hydrogen bond improves the calculated $^{3h}J_{NC'}$ coupling constants and shifts the conformational ensemble sampled from the molecular dynamics (MD) simulations toward the experimentally measured one. Increased charges lead to more narrow distance and angle distributions and improve the agreement between calculated and measured $^{3h}J_{NC'}$ couplings. However, different secondary structures are better represented by different magnitudes of electrostatic interactions—different atomic partial charges in the present work—as indicated by root-mean square deviations (rsmds) between observed and calculated coupling constants $^{3h}J_{NC'}$. The parametrization of the empirical formula is found to be meaningful and robust, but the parameter values are not universal across different proteins and different secondary structural elements ($\alpha$-helices, $\beta$-sheets and loops). Using standard and slightly increased CHARMM charges, predictions for the as-yet unknown scalar coupling constants for the V54A and I6A mutants of protein G are made.

## 1. Introduction

Hydrogen bonds (H-bonds) are important for the stabilization of biomolecular structures, formation of secondary and tertiary structures in proteins, ligand binding, and specificity. Therefore a detailed understanding of the energetics and dynamics of H-bonds is desirable. The ability to measure scalar coupling constants across H-bonds ($^{3h}J_{NC'}$) by using NMR techniques provided valuable information on the overall protein fold.[1] The donor and acceptor atom can now be identified unambiguously and the observed $^{3h}J_{NC'}$ coupling value is a measure of the orbital overlap. In general, scalar couplings are in the range of 0 to −1 Hz. On the basis of a linear combination of atomic orbitals, a purely geometrical model involving distances and angles to characterize an H-bond was proposed to calculate $^{3h}J_{NC'}$ values from structural data.[2] It has been shown that compared to coupling constants from more rigorous and computationally expensive density functional theory (DFT) calculations, deriving $^{3h}J_{NC'}$ values from the H-bond geometry alone is more economical and gives quantitatively correct results. This opens the possibility to calculate scalar $^{3h}J_{NC'}$ coupling constants from molecular dynamics (MD) simulations which has been recently done for different systems.[3,4]

Dynamics is likely to be important to accurately compute $^{3h}J_{NC'}$ couplings as the experimentally observed couplings are averages over the motions up to a time required for coherence transfer (typically 0.1 to 1 s). Compared to computations based on the experimental X-ray or NMR structures alone, significant improvement was found when the conformational motion is included.[3] Furthermore, the

* Corresponding author. Phone/Fax Number: +41 61 267 38 21/ +41 61 267 38 55. E-mail address: m.meuwly@unibas.ch.

10.1021/ct800241d CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/25/2008

analysis showed that the starting structure, in particular the resolution of the experimental structure, is important for the accuracy with which the coupling constants can be calculated.[4] Low-resolution structures can be improved by relaxing the structure, mostly the donor acceptor distance, where the values adapt to the ones found in high-resolution structures. In calculations on high-resolution structures of ubiquitin, protein G, and the Tudor domain for which extensive measurements are available, the best agreement between calculated and observed $^{3h}J_{NC'}$ couplings in a root-mean square (rms) sense is $\approx$0.14 Hz.[4] In addition, by comparing with previous simulations carried out in implicit solvent,[3] it was found that an atomistic representation of the solvent, which is water in the present case, is instrumental for meaningful simulations.

A different approach to model the structural features of H-bonds is to include constraints into the simulations by adding external biasing potentials, e.g. on $^{3h}J_{NC'}$, nuclear Overhauser effects (NOEs), or $S^2$ values, to the force field.[5] Such a procedure lowered the root-mean square diameter (rmsd) between calculated and measured $^{3h}J_{NC'}$ to 0.06 Hz and lead to a shift in the conformational ensemble, such that the distance between donor and acceptor, $r_{H\cdots O'}$, decreases and the directionality of the H-bond, $\theta_1(N-H\cdots O')$, was found to prefer a near-linear (180°) geometry. Such a procedure serves to better characterize the ensemble of protein structures represented by the measurements.

Given the relationship between geometry and scalar coupling constants, it is of interest to explore the influence of force field parameters (in particular nuclear charges) and the empirical parametrization of the model used to calculate $^{3h}J_{NC'}$ values.[2] Usually, force fields are parametrized in view of structural, (optical) spectroscopic, and thermodynamic measurements.[6] The former two provide information on bond lengths and force constants to which the force field can be fitted whereas the latter is rather more sensitive to the nonbonded interactions. However, scalar coupling constants across hydrogen bonds can be expected to be sensitive to parameters describing the interaction strength of a H-bond which is primarily reflected in the nuclear charges or the polarity of the hydrogen bond.

Furthermore, using atomistic simulations, it is possible to consider and separately analyze hydrogen bonds within different secondary structural elements (SSEs, α-helix, $\beta$-sheet, and loop) for which scalar coupling constants can be measured. To this end, three experimentally well-characterized proteins (ubiquitin, protein G, and the Tudor domain), for which an extensive set of $^{3h}J_{NC'}$ values has been measured, is used. Here, we address the following questions: What is the influence of the nuclear charges for atoms involved in the H-bond on the calculated coupling constants and the conformational ensemble sampled from MD simulations? How universal are the parameters in eq 2 (see below) for different proteins? And, are optimal parameter values identical for all secondary structural elements?

The present work is structured as follows. First, the proteins and methods used are described. Next, results for $^{3h}J_{NC'}$ from static structures and from MD simulations are presented and their relationship to the conformational space sampled is explored. Finally, the results are discussed, and predictions for particular mutants of protein G are made which can serve as a test to validate the computational approach used to analyze previously measured scalar coupling constants.

## 2. Materials and Methods

**2.1. Setup of the Proteins.** All molecular dynamics (MD) simulations were carried out with CHARMM[7] and the CHARMM22[6] force field. The starting structures for the three proteins considered, the Tudor domain, protein G, and human ubiquitin, were taken from the Brookhaven Protein Data Bank. All titratable side chains were in their standard protonation state for pH 7. For the Tudor domain, the 1.8 Å X-ray structure (code 1MHN)[8] and 10 NMR structures (code 1G5V)[9] are used. 56 residues are contained in both structures and correspond to residues 90−145. Protein G was set up from the 1.92 Å resolution X-ray structure (code 1PGB)[10] and from the 1.1 Å resolution X-ray structures (codes 1IGD[11] and 2IGD[12]). It consists of 56 residues, and the following mutations were introduced to match the sequence of the protein used in the experimental $^{3h}J_{NC'}$ measurement:[13] T1M, T2Q, V6I, I7L, K19E, E24A A29V, V42E, and, the *N*-terminal was deleted. The mutations were done by replacing the specific sidechains. The 76 residue protein human ubiquitin was taken from the 1.8 Å resolution X-ray structure (code 1UBQ)[14] and ten NMR structures (code 1D3Z).[15] After adding hydrogen atoms, the structures were relaxed with 3000 steps of steepest descent (SD) minimizations. All structures were solvated in a rectangular box of pre-equilibrated TIP3P water (box sizes are 50 × 40 × 40, 56 × 47 × 40, and 65 × 50 × 47 Å$^3$, respectively, for Tudor domain, protein G, and ubiquitin). After removing solvent molecules which overlap with the protein, the systems contained between 8000 and 15 000 atoms.

**2.2. Molecular Dynamics Simulations.** Molecular dynamics simulations were carried out with periodic boundary conditions and images were updated every 10 time steps. A 12 Å cutoff was applied to the shifted electrostatic and switched van der Waals interactions. The systems have an overall charge of −3 for the Tudor domain, −4 for protein G, and 0 for ubiquitin. No counterions were added to neutralize the systems. The systems were heated and equilibrated at 300 K for 50 ps. All MD simulations were carried out using the leap Verlet algorithm with a time step of 1 fs and constraining hydrogen atoms with SHAKE,[16] which is consistent with the way in which force field parameters are optimized.[6] Several trajectories from different starting structures were run to sample the conformational ensemble. Overall, the analysis of the H-bond dynamics is based on 19 ns for the Tudor domain, 10 ns for protein G, and 15 ns for ubiquitin.

*Atomic Partial Charges in the H-Bond Motif.* To study the influence of varying atomic partial charges on the conformational ensemble, MD simulations were carried out using identical protocols. The atomic charges on the nitrogen, hydrogen, and oxygen atoms involved in the H-bonds (for the definition of the H-bonding motif, see Figure 1) were

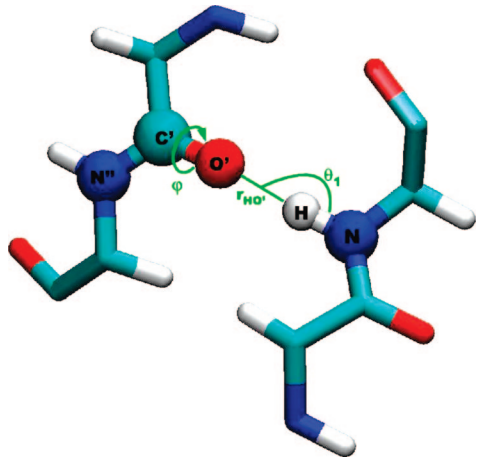



***Figure 1.*** Ball and stick representation of the H-bond motif N−H···O′=C′−N″ together with the geometric properties incorporated in eq 1 to calculate $^{3h}J_{NC'}$ values: the donor−acceptor distance H···O′ ($r_{H\cdots O'}$), the directionality angle N−H···O′ ($\theta_1$), and the dihedral angle H···O′=C′−N″ ($\phi$).

***Table 1.*** Atomic Partial Charges for H-Bond Atoms Applied in MD Simulations and ab initio Charges for Specific H-Bond Sites

| | partial charges [e] | | | | sum |
|---|---|---|---|---|---|
| | N−H···O=C | | | | |
| $q_{-0.04}$ | −0.45 | 0.27 | −0.49 | 0.51 | −0.16 |
| $q_0$ | −0.47 | 0.31 | −0.51 | 0.51 | −0.16 |
| $q_{+0.04}$ | −0.49 | 0.35 | −0.53 | 0.51 | −0.16 |
| $q_{+0.1}$ | −0.52 | 0.41 | −0.56 | 0.51 | −0.16 |
| $q_{+0.2}$ | −0.57 | 0.51 | −0.61 | 0.51 | −0.16 |
| $q_{+0.4}$ | −0.67 | 0.71 | −0.71 | 0.51 | −0.16 |
| J12[a] | −0.64 | 0.43 | −0.67 | 0.70 | −0.18 |
| J12[b] | −0.59 | 0.40 | −0.62 | 0.65 | −0.16 |
| J19[a] | −0.69 | 0.42 | −0.63 | 0.70 | −0.20 |
| J19[b] | −0.55 | 0.34 | −0.51 | 0.56 | −0.16 |

[a] ab initio NBO charges for coupling no. J12 and J19 of ubiquitin as obtained. [b] Scaled NBO charges to maintain the sum −0.16 e.

systematically increased and decreased. For all models, the total charge of the H-bond motif was preserved ($q_O + q_H + q_N = -0.16$ e). The different charge sets (see Table 1) are labeled $q_0$ for standard CHARMM charges and $q_{-0.04}$, $q_{+0.04}$, $q_{+0.1}$, $q_{+0.2}$, and $q_{+0.4}$ according to the charge increase/decrease on the hydrogen atom.

The variation of the partial charges is motivated from results of ab initio electronic structure calculations on two different H-bond motifs of ubiquitin (couplings J12 and J19) at the B3LYP/6-31G* level of theory. To assess possible electronic coupling between neighboring residues, the ab initio calculations included the preceding and succeeding amino acid of H-bond donor and acceptor (see Figures S1 and S2 in the Supporting Information). In the electronic structure calculations, coordinates for all except the H-bond hydrogen atom were kept fixed. These calculations were carried out with the Gaussian suite of programs.[17] Charges from a natural bond orbital (NBO) analysis are summarized in Table 1. The total charge of the H-bonding motif slightly differs from −0.16 e which is the value from the CHARMM force field used here. Therefore, appropriately scaled charges to make them directly comparable to the CHARMM force field are also reported.

**2.3. Analysis of the Results.** From the structures and the MD simulations, geometries describing the hydrogen bonds (the distances H···O′ ($r_{H\cdots O'}$) and N···O′ ($r_{N\cdots O'}$), the angles N−H···O′ ($\theta_1$) and H···O′=C′ ($\theta_2$), and the dihedral H···O′=C′−N″ ($\phi$)) were extracted every 0.1 ps. This is done for all atoms for which scalar couplings across H-bonds are observed experimentally except for coupling no. 1 of ubiquitin, because for this coupling homotopic hydrogen atoms exist. A total of 15, 32, and 29 *J*-couplings were analyzed for the Tudor domain, protein G, and ubiquitin, respectively (see Figure 2). These H-bonds are located in different SSEs, namely the α-helix and β-sheet and loop regions (see Tables S6−S8 in the Supporting Information).

Using a model based on a linear combination of atomic orbitals, a formula which relates geometries characterizing a hydrogen bond to scalar $^{3h}J_{NC'}$ coupling constants was proposed, parametrized, and tested:[2]

$$^{3h}J_{NC'}(r_{HO'}, \theta_1, \phi) = [\alpha \cos^2 \theta_1 + f(\phi)]e^{\beta(r_{HO'} - r^0_{HO'})} - 0.10 \text{ Hz} \quad (1)$$

The coordinates involved are the H···O′ distance $r_{HO'}$, the N−H···O′ angle $\theta_1$, and the H···O′=C′−N″ dihedral $\phi$ (see Figure 1 for a definition of coordinates). The contribution of the dihedral term $f(\phi)$ was found to be small and primarily removes systematic differences between α-helices and β-sheets.[2,4] Thus, eq 1 was simplified to

$$^{3h}J_{NC'}(r_{H\cdots O'}, \theta_1) = \alpha \cos^2(\theta_1)e^{\beta r_{H\cdots O'}} + \delta \text{ Hz} \quad (2)$$

which is the form used in the present work. The original parameters in eq 2 were fitted to 34 experimentally determined $^{15}N-^{13}C'$ coupling constants in protein G and yielded the following values: $\alpha_B = -357$ Hz, $\beta_B = -3.20$ Å$^{-1}$, and $\delta_B = 0$ Hz, where the index "B" refers to Barfield's work. The distance $r^0_{HO'}$ was 1.76 Å, the smallest H···O′ distance found in crystallographic structures.[2]

One of the aims of the present work is to study the universality of eq 2 for the Tudor domain, protein G, and ubiquitin. For this, the coupling constant $^{3h}J_{NC'}$ was calculated using eq 2 with the original parameters $\alpha_B$ and $\beta_B$ for each snapshot from the MD simulations and then averaged over all structures from which the ensemble averaged coupling constant, $\langle ^{3h}J_{NC'}\rangle$ was obtained. Furthermore, optimal parameters $\alpha^{opt}$ and $\beta^{opt}$ which minimize the sum of squares $\sum(^{3h}J^{calc}_{NC'} - {}^{3h}J^{obs}_{NC'})^2$ were determined where the sum extends over all couplings. As the parametrization of eq 2 suggests that the parameters α and β are correlated; α, which is only a scaling factor, is kept fixed, whereas β is optimized. The optimization is carried out for all $^{3h}J_{NC'}$ couplings and individually for couplings located in particular SSEs. Tests whereby the parameter α was varied were carried out after an optimal value for β had been found. However, the rmsd could usually not be improved any further. Thus, $\alpha_B = -357$ Hz was used throughout the present work.

## 3. Results

In the following, results for the Tudor domain, protein G, and ubiquitin are discussed. The Tudor domain contains

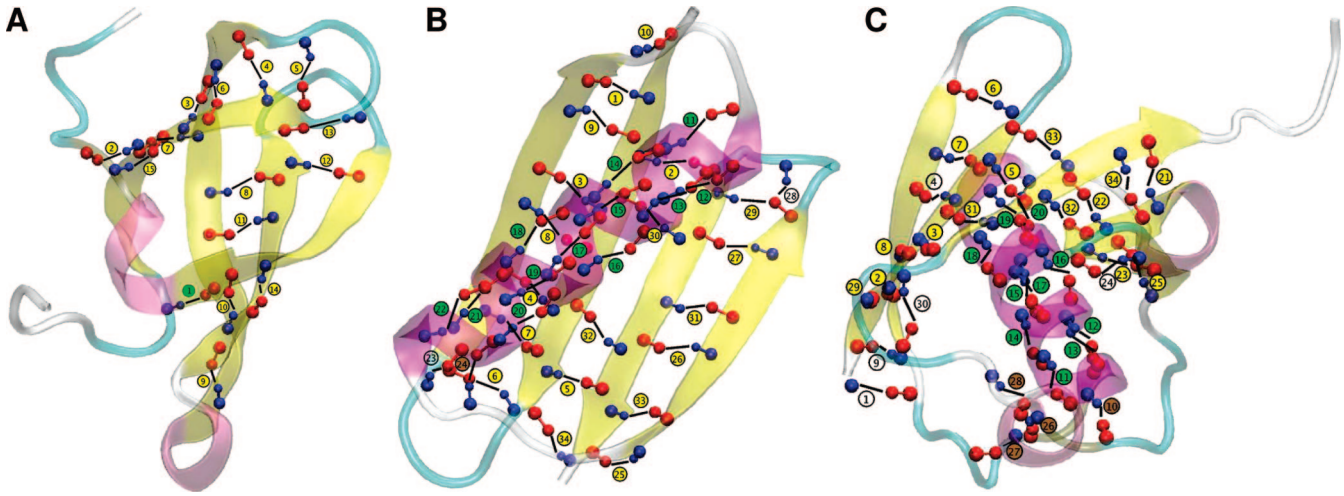



***Figure 2.*** $^{3h}J_{NC'}$ coupling sites. The structure of (A) the Tudor domain, (B) protein G, and (C) ubiquitin in cartoon representation to highlight the different secondary structures. α-Helices are colored pink, β-strands are in yellow, and loops are in cyan. Hydrogen bonds corresponding to the experimentally measured scalar couplings are depicted in ball and stick representation for N—H (blue) and O═C (red). Each coupling is enumerated (J1−J*n*). Coupling constants used in this work are colored according to their secondary structure location in the protein: α-helix (green) and β-sheet (yellow) and loop (brown).

***Table 2.*** Average rmsds between Calculated and Experimentally Measured Scalar Couplings $^{3h}J_{NC'}$ for Static Structures (X-ray and NMR)[a]

| | all | | α-helix | | β-sheet | | loop | |
|---|---|---|---|---|---|---|---|---|
| protein (coupl) | $\beta$ [Å$^{-1}$] | rmsd [Hz] | $\beta$ [Å$^{-1}$] | rmsd [Hz] | $\beta$ [Å$^{-1}$] | rmsd [Hz] | $\beta$ [Å$^{-1}$] | rmsd [Hz] |
| Tudor (15) | −3.20 | 0.25 | | | −3.20 | 0.25 | | |
| fitted $\beta$ | −3.05 | 0.20 | | | −3.05 | 0.20 | | |
| protein G (32) | −3.20 | 0.14 | −3.20 | 0.12 | −3.20 | 0.15 | | |
| fitted $\beta$ | −3.28 | 0.12 | −3.26 | 0.09 | −3.25 | 0.13 | | |
| ubiquitin (29) | −3.20 | 0.31 | −3.20 | 0.20 | −3.20 | 0.40 | −3.20 | 0.14 |
| fitted $\beta$ | −3.46 | 0.23 | −3.38 | 0.10 | −3.42 | 0.28 | −3.05 | 0.13 |

[a] Results are provided for both $\beta_B$ and optimized $\beta^{opt}$ to minimize the rmsd between calculated and observed couplings.

hydrogen bonds in β-sheets; for protein G, they are located in α-helices and β-sheets; ubiquitin has H-bonds in α-helices and β-sheets and loop regions. Results are reported and discussed separately for static structures and those from MD simulations.

**3.1. Static Structures.** First, eq 2 with parameters $\alpha_B$ and $\beta_B$ was assessed by calculating rmsds between observed and calculated scalar couplings for the static X-ray and NMR structures of all three proteins. Including all experimentally observed couplings, the rmsds for the Tudor domain, protein G, and ubiquitin are 0.25, 0.14, and 0.31 Hz, respectively (see Table 2). The agreement between calculation and experiment is comparable to previous results.[2,4] Considering scalar coupling constants in particular SSEs, it is found that rmsds for α-helices (for ubiquitin 0.20 compared to 0.31 Hz) and loops (for ubiquitin 0.14 compared to 0.31 Hz) are smaller than by calculating rmsds including all couplings (see Table 2). In contrast, β-sheet couplings are less accurately estimated (for ubiquitin 0.40 compared to 0.31 Hz). This suggests that it is of interest to also consider individual parametrizations $\beta_H$ (for α-helices), $\beta_S$ (for β-sheets), and $\beta_L$ (for loop regions) for different SSEs of eq 2.

Averages $\langle^{3h}J_{NC'}\rangle$ of the experimentally measured coupling constants for individual SSEs of each protein showed significant differences (up to 0.25 Hz in ubiquitin; see Table 3) in the magnitude of the scalar couplings. More negative values for the scalar couplings are found for β-sheets and less negative values for loop regions and α-helices (see Table 3), i.e. β-sheet hydrogen bonds are stronger than the ones in α-helix or loops. Thus, for all static protein structures, the parameter $\beta$ was individually optimized for the three structure elements (see Table 2). It is found that these optimized $\beta^{opt}$ values can differ considerably from $\beta_B = -3.20$ Å$^{-1}$. For protein G, the optimal $\beta$ values for α-helix and β-sheet are similar, −3.26 and −3.25 Å$^{-1}$, respectively. However, for ubiquitin, $\beta$ differs significantly for each secondary structural element: −3.38 Å$^{-1}$ for α-helix, −3.42 Å$^{-1}$ for β-sheet, and −3.05 Å$^{-1}$ for loops. The parameter optimization decreased the rmsds for all secondary structure elements, although not equally well. Again, α-helices and loops are found to perform better than β-sheet (protein G 0.09 Hz rmsd (α-helix) vs 0.13 Hz (β-sheet), ubiquitin 0.10 Hz rmsd (α-helix) vs 0.28 Hz (β-sheet)).

***Table 3.*** Average Experimentally Measured $^{3h}J_{NC'}$ Scalar Couplings for Different Secondary Structure Elements

| | $^{3h}J_{NC'}$ [Hz] | | |
|---|---|---|---|
| | ubiquitin | protein G | Tudor domain |
| overall | −0.50 ± 0.171 | −0.43 ± 0.179 | −0.55 ± 0.104 |
| helix | −0.35 ± 0.149 | −0.37 ± 0.211 | |
| sheet | −0.61 ± 0.094 | −0.47 ± 0.153 | −0.55 ± 0.104 |
| loop | −0.43 ± 0.161 | | |

***Table 4.*** $\beta^{opt}$ Parameters [Å$^{-1}$ ] with Corresponding $^{3h}J_{NC'}$ rmsd [Hz] for Each Charge Set and Secondary Structure Element Calculated from Averages over MD Simulations of All Protein Systems Considered[a]

| charge set | overall | | helix | | sheet | | loop | |
|---|---|---|---|---|---|---|---|---|
| | rmsd | $\beta$ | rmsd | $\beta$ | rmsd | $\beta$ | rmsd | $\beta$ |
| $q_{-0.04}$ | 0.21 | −3.09 | 0.21 | −3.04 | 0.21 | −3.09 | 0.21 | −3.03 |
| $q_0$ | 0.16 | −3.19 | 0.15 | −3.20 | 0.14 | −3.19 | 0.20 | −3.09 |
| $q_{+0.04}$ | **0.15** | −3.30 | **0.14** | −3.32 | 0.14 | −3.30 | 0.21 | −3.20 |
| $q_{+0.1}$ | **0.15** | −3.45 | 0.15 | −3.35 | **0.13** | −3.45 | 0.17 | −3.34 |
| $q_{+0.2}$ | **0.15** | −3.67 | 0.18 | −3.82 | 0.17 | −3.62 | **0.11** | −3.54 |
| $q_{mixed}$ | 0.22 | −3.31 | 0.18 | −3.32 | 0.25 | −3.45 | 0.23 | −3.34 |

[a] Optimal charge sets for specific secondary structures are emphasized.

**3.2. Simulations: Influence of Modified Force Field Parameters.** NMR experiments, from which scalar coupling constants $^{3h}J_{NC'}$ are determined, are by their very nature sensitive to the dynamics of this system. Therefore, MD simulations using explicit solvent were carried out to calculate ensemble-averaged $^{3h}J_{NC'}$ couplings. To better sample the conformational subspace, different starting structures were used. Snapshots, separated by 0.1 ps, along the trajectories with total lengths of 19, 15, and 10 ns for Tudor domain, ubiquitin, and protein G, respectively, were analyzed in the same way as the static structures described earlier, and an ensemble average $\langle^{3h}J_{NC'}\rangle$ was calculated. In addition to the static structures, different atomic partial charge sets were employed to investigate the influence of modified electrostatic interactions on the H-bond geometries and the conformationally averaged $^{3h}J_{NC'}$ couplings. The effect of the atomic partial charges is assessed by considering the rmsd between calculated and experimental $^{3h}J_{NC'}$ data and the ensemble distributions of the H···O′ distance, $r_{H\cdots O'}$, and the N−H···O′ angle, $\theta_1$.

*Quality of $^{3h}J_{NC'}$ Coupling Prediction.* The relationship between partial charges and the rmsd between observed and calculated scalar coupling constants is tested by including all couplings ($\sum_{all}(^{3h}J_{NC'}^{calc} - {}^{3h}J_{NC'}^{obs})^2$) and distinguishing between SSEs ($\sum_{SSE}(^{3h}J_{NC'}^{calc} - {}^{3h}J_{NC'}^{obs})^2$). Thus for each charge set the optimal $\beta$, $\beta^{opt}$ was determined. The optimized values $\beta^{opt}$ and corresponding rmsds are summarized in Tables 4 and 5.

It is found that different charge sets affect the calculated coupling constants and also the value $\beta^{opt}$ which minimizes $\sum(^{3h}J_{NC'}^{calc} - {}^{3h}J_{NC'}^{obs})^2$. Reducing the charge on the hydrogen atom by 0.04 e leads to a larger rmsd for all couplings and for couplings in particular SSEs compared to the standard charges $q_0$. Increasing the partial charge on the H-atom initially improves the agreement between calculated and experimentally measured $^{3h}J_{NC'}$ couplings. Depending on whether all couplings or couplings in particular SSEs are considered, the optimal charge set varies, as do the parameters $\beta^{opt}$. Considering all couplings, the lowest rmsd is found for charge set $q_{+0.1}$ (see the red line in Figure 3). Calculations for the three proteins individually yielded different optimal charge sets, namely $q_{+0.1}$ for the Tudor domain, $q_0$ for protein G, and $q_{+0.2}$ for ubiquitin (see Table 5). Thus, the ideal charge set depends on the particular protein. More generally, it is found that increased charges perform better. Comparison of rmsds between calculated and observed $\langle^{3h}J_{NC'}\rangle$ obtained from static structures with the ones from MD simulations showed that dynamics, i.e. conformational sampling in explicit solvent, leads to improved $\langle^{3h}J_{NC'}\rangle$ values (see Tables 2 and 5). For the Tudor domain and ubiquitin, an appreciable improvement is found. Including dynamics decreases the rmsd by 0.08 and 0.11 Hz, respectively. By contrast for protein G, $^{3h}J_{NC'}$ coupling predictions from static structures and from simulations are quite similar. This is consistent with previous simulations using a different protocol (see Discussion and Conclusions).[4]

It is also of interest to consider individual SSEs and to determine optimal charge sets and best achievable rmsds between experiment and simulations. Different SSEs favor distinct partial charges and also have significantly different parameter values ($\beta_H^{opt}$ for α-helix, $\beta_S^{opt}$ for $\beta$-sheet, and $\beta_L^{opt}$ for loop structures). Analysis of the entire data from all three proteins in view of the SSEs yielded the following optimal charge sets: slightly enhanced charges $q_{+0.04}$ for α-helix and higher charges for $\beta$-sheet ($q_{+0.1}$) and loop regions ($q_{+0.2}$), respectively (see Figure 3 and Table 4). The corresponding $\beta^{opt}$ values are $\beta_H^{opt} = -3.32$ Å$^{-1}$, $\beta_S^{opt} = -3.45$ Å$^{-1}$, and $\beta_L^{opt} = -3.54$Å$^{-1}$. Detailed analysis of SSEs for each protein showed that the lowering of the rmsd for the Tudor domain and ubiquitin observed from MD simulations is mainly due to improved couplings within the $\beta$-sheet structures.

On the basis of the result that different partial charges are favored by different SSEs, an additional 15 ns simulation using a mixed charge set, $q_{mixed}$, was run for ubiquitin. This charge set consisted of $q_{+0.04}$ for H-bonds located in α-helix structures and loops and $q_{+0.1}$ for $\beta$-sheet couplings. The use of $q_{mixed}$ showed no obvious distortions of the protein structure as judged from the H-bond geometries ($r_{H\cdots O'}$ and $\theta_1$) sampled during the simulation (not shown). In a first step, the $^{3h}J_{NC'}$ couplings using $q_{mixed}$ were estimated by using optimal values for $\beta$ determined above: $\beta_H = -3.32$ Å$^{-1}$, $\beta_S = -3.45$ Å$^{-1}$, and $\beta_L = -3.54$ Å$^{-1}$. It was found that the mixed charges, $q_{mixed}$, lead to inferior rmsds. Most notably, the $\beta$-sheet couplings were reproduced inaccurately, with a rmsd of 0.25 Hz as compared to 0.13 Hz from $\beta$-sheet couplings of all three proteins or even 0.07 Hz obtained with set $q_{+0.4}$ for ubiquitin. Couplings in α-helix and loop structures have comparable values for $q_{mixed}$ and $q_{+0.04}$ charge sets. To determine whether the discrepancy of $\langle^{3h}J_{NC'}\rangle$ estimation is due to the use of the parameter $\beta_S$ fitted for secondary structure and charge set with data from all three proteins, $\beta_S$ was reoptimized for the current structure ensemble. With $\beta^{opt} = -3.22$ for $\beta$-sheets, an improved rmsd of 0.12 Hz was found which is in accord with 0.13 Hz rmsd from the entire data set (compare Tables 4 and 5). Finally, it is also found that increased partial charges across the H-bond not only improve the calculated $\langle^{3h}J_{NC'}\rangle$ values but they also decrease the standard deviations of the calculated couplings. Results for ubiquitin are shown in Figure 4, which are representative for all three proteins studied.

*Geometric Properties of Ensemble.* The influence of different atomic partial charges on the conformational ensemble can also be characterized from the MD simulations and compared with experiment. For all H-bond motifs, the

***Table 5.*** Optimal Charge Set and $\beta^{opt}$ Obtained from MD Simulations Given for Each Protein ($^{3h}J_{NC'}$ rmsd [Hz] and $\beta^{opt}$ [Å$^{-1}$])

| | Tudor | | | protein G | | | ubiquitin | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | charges | $\beta^{opt}$ | rmsd | charges | $\beta^{opt}$ | rmsd | charges | $\beta^{opt}$ | rmsd | charges | $\beta^{opt}$ | rmsd |
| overall | $q_{+0.1}$ | −3.45 | 0.12 | $q_0$ | −3.18 | 0.14 | $q_{+0.2}$ | −3.45 | 0.12 | $q_{mixed}$ | −3.22 | 0.16 |
| helix | | | | $q_0$ | −3.26 | 0.12 | $q_{+0.2}$ | −3.53 | 0.10 | $q_{mixed}$ | −3.24 | 0.18 |
| sheet | $q_{+0.1}$ | −3.45 | 0.12 | $q_0$ | −3.14 | 0.16 | $q_{+0.2}$ | −3.45 | 0.07 | $q_{mixed}$ | −3.22 | 0.12 |
| loop | | | | | | | $q_{+0.2}$ | −3.30 | 0.14 | $q_{mixed}$ | −3.10 | 0.22 |

donor−acceptor distance, $r_{H\cdots O'}$, and the H-bond directionality, $\theta_1$, are considered separately for each protein and each charge set. Figure 5 summarizes the results for the Tudor domain, protein G, and ubiquitin. Experimental values ($r_{H\cdots O'}$, $\theta_1$) from X-ray and NMR structures, where "experiment" refers to optimized H-positions with heavy atoms kept fixed, are indicated as blue squares and triangles, respectively, whereas structures from the simulation ensemble are individual dots (50 snapshots each coupling). The analysis was also carried out for N···O′ distances which are directly available from the experiment and the same effects, as described below, were found (see Figure S3 in the Supporting Information). Variation of the atomic partial charges influences the conformational ensemble, in particular the H-bond length $r_{H\cdots O'}$ and the H-bond angle $\theta_1$ which are in turn used to calculate $^{3h}J_{NC'}$ couplings. Increased partial charges restrict the conformational space to H···O′ distances around 2 Å and favor linear H-bonds, i.e. $\theta_1 \approx 180°$. For increased charges, the geometry distributions from the simulations approach those from the experiments (see Figure 5). Thus, the conformational space sampled from the simulations better reflects the one observed and analyzed experimentally. The tighter geometries of both, H-bond distances and angles, is also reflected in decreased standard deviations for the $^{3h}J_{NC'}$ couplings from simulations with increased charges. This is related to increased Coulomb interactions. The improvement

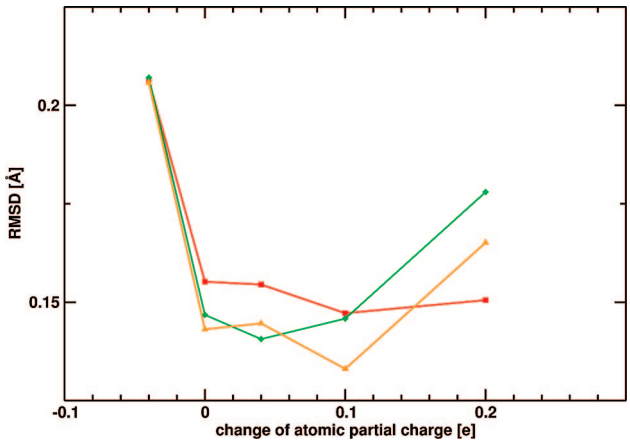



***Figure 3.*** Performance of charge sets. The rmsd between experimental and calculated $\langle^{3h}J_{NC'}\rangle$ couplings is shown as a function of the charge set used for all couplings (red), for α-helix (green), and β-sheet (orange). Larger charge disparity leads to lower rmsd. Charge sets $q_0$ to $q_{+0.1}$ are suitable. In addition, the secondary structures have minor preferences for a specific charge set, i.e. $q_{+0.04}$ for α-helix and $q_{+0.1}$ for β-sheet.

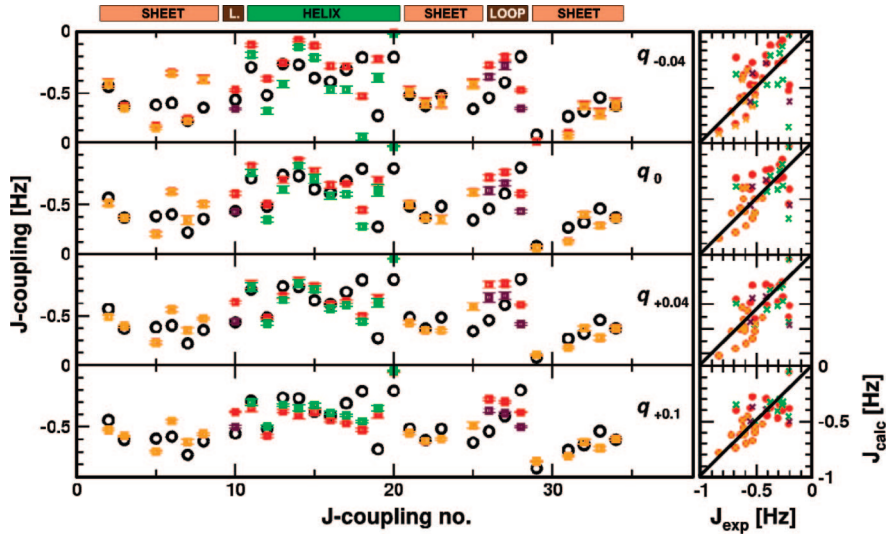



***Figure 4.*** Calculated $^{3h}J_{NC'}$ values for ubiquitin with their respective standard errors from the different MD runs for the different charge sets ($q_{-0.04}$, $q_0$, $q_{+0.04}$, and $q_{+0.1}$) used in the present work. Secondary structure elements are indicated on top and predictions with optimized parameter $\beta^{opt}_{SSE}$ are colored as follows: overall (red), α-helix (green), β-sheet (orange), and loop (brown). For comparison, experimental measurements are indicated as black circles. Correlation between $J_{calc}$ and $J_{exp}$ highlighting the quality of prediction depicted on the right. Note that for increased charge sets, the standard error decreases (left graph) and the agreement between calculated and experimentally measured $^{3h}J_{NC'}$ values is improved (right graph). In particular, parameter optimization for individual secondary structures (crosses) is superior to optimizations including all couplings (red circles). $\beta^{opt}_{SSE}$ correspond to β values optimized for particular SSEs and charge sets; see Table 5.

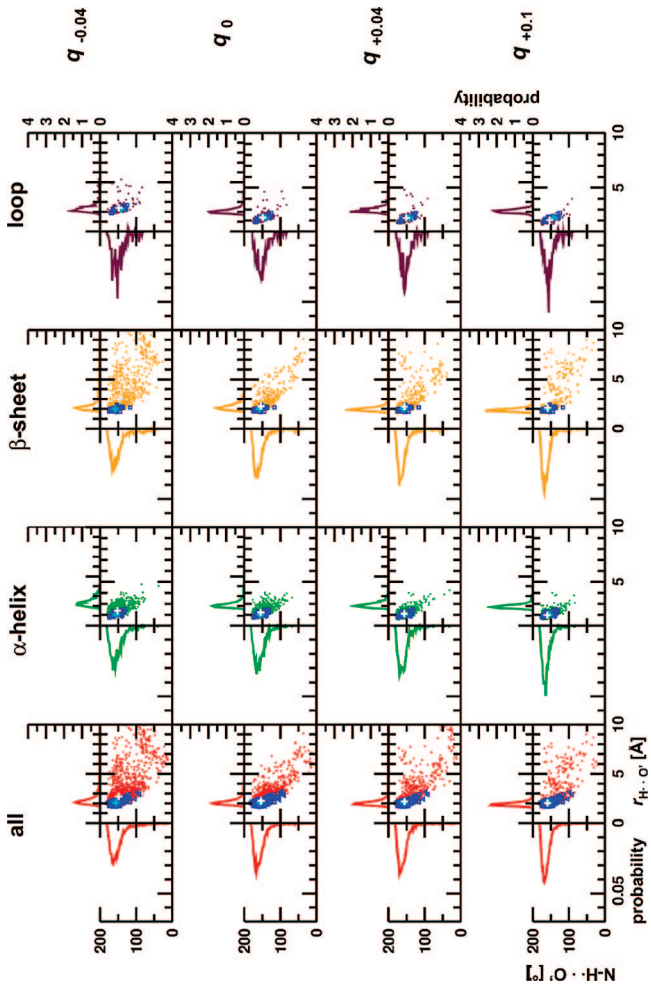



***Figure 5.*** H-Bond geometry ensemble distributions. The geometric H-bond properties $r_{H\cdots O'}$ and $\theta_1$ (angle N−H···O′) of 50 snapshots are shown as points for every coupling of the three proteins studied. Values found experimentally, i.e. H-position optimization while keeping heavy atom positions frozen, are indicated in blue (squares X-ray, triangles NMR). Experimental (cyan) and MD (gray) centers of gravity are shown as a cross (see Table 6). In addition, the probability distributions of the donor−acceptor distance and angle are given. Charge sets are arranged in rows and secondary structures are in columns.

of the H-bond geometries is not uniform across the different SSEs; i.e. they have different sensitivities with respect to atomic partial charges. For protein G, H-bonds located in the α-helix have a narrow $r_{H\cdots O'}$ and $\theta_1$ distribution, close to the experimental values, which is largely independent of the charge set used. Disrupted H-bonds, characterized by a large H···O′ separation, mainly occur at the end of $\beta$-sheets. Even with an exaggerated charge difference ($q_{+0.4}$) H-bonds with very large donor−acceptor distances ($r_{H\cdots O'}$ > 5.0 Å) are observed for coupling J5 of the Tudor domain and couplings J1, J6, J7, J10, J25, and J34 of protein G.

In summary, increased atomic partial charges shift the geometric ensemble toward the one found experimentally without affecting the overall dynamics (see discussion). Furthermore, the degree to which this occurs depends on the particular SSE considered. Finally, which charge set is suited to best reproduce on average (in a rmsd sense) experimentally measured $^{3h}J_{NC'}$ couplings depends on the protein studied: they are $q_0$ (i.e., CHARMM22 in the present case) for protein G, $q_{+0.1}$ for the Tudor domain, and $q_{+0.2}$ for ubiquitin.

**3.3. Influence of H-Bond Location in Protein Structure.** Although increased partial charges influence the conformational ensemble and improve the overall agreement with experiment, outliers occur. To identify outliers an arbitrary threshold of $r_{H\cdots O'}$ > 5.0 Å is used. Such a value corresponds to an inexistent H-bond or a vanishing scalar coupling. As stated previously, hydrogen bonds corresponding to outliers are in general located at the end of $\beta$-sheet structures, e.g. adjacent to loop regions or chain termini. Throughout all charge sets used here no outlier (out of 29) for ubiquitin and only two (out of 15) for the Tudor domain (couplings 5 and 13) were identified. Contrary to that, for protein G, 9 couplings (out of 32) have donor−acceptor distances larger than 5.0 Å. Half of them disappear if charge sets $q_{\geq 0}$ are used. Persisting outliers in any of the increased charge sets are couplings J1, J6, J7, J10, J25, and J34. Identifying the optimal charge set for individual couplings showed that some $^{3h}J_{NC'}$ couplings are best captured with the $q_{-0.04}$ charge set, although in general stronger Coulombic interactions are superior. These are for the Tudor domain coupling no. J13, for protein G nos. J8, J13, J15, J19, J20, J27, and J33, and for ubiquitin nos. J2, J7, J23, and J33. However, their location includes α-helix and $\beta$-sheet structures where no systematic pattern is observed. In contrast, loop region H-bonds seem to require at least charge set $q_0$.

It is also of interest to consider whether the degree of solvent exposure of the residues involved in the H-bonds can be correlated with the quality of the calculated $J$-couplings. For this, the fraction of the solvent accessible surface area (SASA) to the total residue surface is calculated for each residue. A threshold of SASA >30% is typically used to define solvent exposure. Residues are considered deeply buried ("core" residues) if they have a SASA of <2%. The majority of experimentally measured H-bonds are located in the protein's interior (SASA < 30%). The few solvent-exposed couplings are in loop and α-helices. Couplings with appreciable errors (arbitrarily taken as $|^{3h}J_{NC'}^{exp} - {}^{3h}J_{NC'}^{calc}|$ > 0.1 Hz) are found both in the "core" and in the surface area of the proteins. In addition, no apparent correlation between the polarity of the H-bond and the degree of solvent exposition of the residues involved in a H-bond was observed.

## 4. Discussion and Conclusions

In the present study, the applicability, parametrization, generality, transferability, and susceptibility to changes in the force field parameters of an empirical formula that relates geometric properties of a H-bond to measured scalar coupling constants from NMR experiments was investigated in detail. Analysis of optimized structures (starting from X-ray and NMR structures) and extensive MD simulations in explicit solvent showed that overall standard charges of the CHARMM22 force field ($q_0$) perform reasonably well for estimating $^{3h}J_{NC'}$ couplings with the original $\beta_B$ previously derived from structures and DFT calculations alone. This is

***Table 6.*** Geometric Centers of Gravity for Experiment and Simulations

| | overall | | helix | | sheet | | loop | |
|---|---|---|---|---|---|---|---|---|
| | $r_{H\cdot\cdot O'}$ [Å] | $\theta_1$ [deg] | $r_{H\cdot\cdot O'}$ [Å] | $\theta_1$ [deg] | $r_{H\cdot\cdot O'}$ [Å] | $\theta_1$ [deg] | $r_{H\cdot\cdot O'}$ [Å] | $\theta_1$ [deg] |
| $q_{-0.04}$ | 2.69 | 148.10 | 2.37 | 151.58 | 3.01 | 145.98 | 2.40 | 147.80 |
| $q_0$ | 2.26 | 153.37 | 2.22 | 154.55 | 2.27 | 154.16 | 2.28 | 149.59 |
| $q_{-0.04}$ | 2.21 | 155.60 | 2.11 | 156.96 | 2.23 | 156.17 | 2.22 | 151.23 |
| $q_{+0.01}$ | 2.03 | 159.51 | 1.95 | 160.32 | 2.11 | 159.40 | 2.05 | 153.78 |
| experiment | 2.07 | 156.65 | 2.01 | 155.47 | 1.98 | 158.67 | 2.25 | 141.90 |

different if the conformational ensemble is considered where larger polarity of the H-bond shifts the distribution toward the experimentally measured one (see Table 6).

By allowing the parameter $\beta$ to vary, the rmsd between calculated and experimentally determined $^{3h}J_{NC'}$ couplings can be reduced from 0.25 to 0.20 (Tudor domain), 0.14 to 0.12 (protein G), and 0.31 to 0.23 (ubiquitin), respectively. For couplings in particular SSEs, this can lead to improvements in rmsd of up to 50% (see Table 2). In all cases, $\beta^{opt}$ differs appreciably from $\beta_B = -3.20$ Å$^{-1}$. It is also found that couplings in different SSEs are characterized by distinct values of $\beta$. Using independent multinanosecond MD simulations starting from various experimental X-ray and NMR structures, it was found that typically larger polarity of the H-bond (stronger interaction) leads to improved estimates for calculated $^{3h}J_{NC'}$ couplings. In all cases, the rmsd can be reduced to between 0.12 and 0.14 with the largest improvement for ubiquitin and the smallest one for protein G. Improvements for couplings located in particular SSEs can be even larger, such as for couplings in $\beta-$ sheets in ubiquitin which have a rmsd of 0.07 with increased polarity (from the $q_{+0.2}$ charge set). Another finding of the present work is that, considering all three proteins studied here, the lowest rmsd in α-helices and $\beta$-sheets are combinations (($q_{+0.04}$, $\beta = -3.32$ Å$^{-1}$) and ($q_{+0.10}$, $\beta = -3.45$ Å$^{-1}$)), respectively. It is worthwhile to point out that the rmsds from using the "standard" parameters ($q_0$, $\beta = -3.20$ Å$^{-1}$) in both cases are only slightly worse. However, significant differences are found for the distribution of the conformational ensemble which hint toward a substantial improvement if larger charges are used (see Figure 5). It is instructive to compare "centers of gravity"—$r_{H\cdots O'}$ and $\theta_1$ averages—for all hydrogen bonds in different SSEs from experiment and simulations (see Table 6). As can be expected, charge variation has a more pronounced effect on the H···O′ distance compared to the N−H···O′ angle. Increased polarity ($q_{+0.04}$ and $q_{+0.1}$) captures the experimentally measured average $r_{H\cdots O'}$ distance even though H-bond linearity is somewhat overestimated. Application of different atomic charges does not distort the protein structure which is confirmed by rmsd values below 2 Å compared to the starting structure (see Table S5 in the Supporting Information). In addition, calculated *B*-factors, which characterize flexible and rigid parts in the protein, agree quite well with experiment independent of the charges used (see Figure S4 in the Supporting Information). Together these results underline the applicability of modified atomic charges without altering the protein dynamics. As mentioned in the introduction, the expectation that modified partial charges across the H-bond lead to better scalar coupling constants rests on the observation that the geometry and energetics of a hydrogen bond is primarily described by electrostatic interactions in a force field. However, it might also be of interest to consider modified van der Waals parameters and their effect on calculated scalar coupling constants.

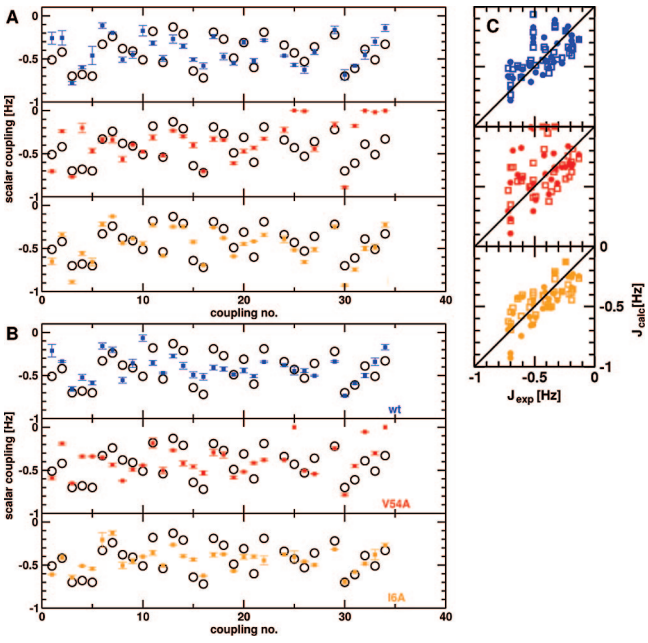



***Figure 6.*** Coupling constants for protein G mutants V54A and I6A involved in coupling. For charge sets $q_0$ (A) and $q_{+0.1}$ (B), the calculated $^{3h}J_{NC'}$ values are shown for protein G wildtype (blue) and for the J32 mutants V54A (red) and I6A (orange). Experimental data for the wild type protein are shown as black circles. (C) Correlation between $J_{calc}$ and $J_{exp}$ for charge set $q_0$ (filled circle) and $q_{+0.1}$ (open square).

It is also of interest to compare the results from the present simulations with previous work. In a recent study, the same three proteins were studied with the same force field (CHARMM22) using NAMD and particle mesh Ewald (PME) for treating the long-range electrostatics.[4] This study was mainly concerned with assessing the differences between simulations in implicit and explicit solvent and with the question whether carrying out MD simulations (trajectory length between 0.5 and 1.4 ns) improves the conformational ensemble as reflected by the measured $^{3h}J_{NC'}$ coupling constants. For the static structures (see Table 1 and ref 4), rmsds were 0.15, 0.20, and 0.23 for protein G (structure 2IGD), ubiquitin (1UBQ), and the Tudor domain (1MHN), respectively. This compares with 0.14, 0.31, and 0.25 from the present work which, however, are averages over several structures (see section 3.1). Including dynamics, the earlier study found rmsds of 0.14, 0.15, and 0.30 (0.15 for rescaled

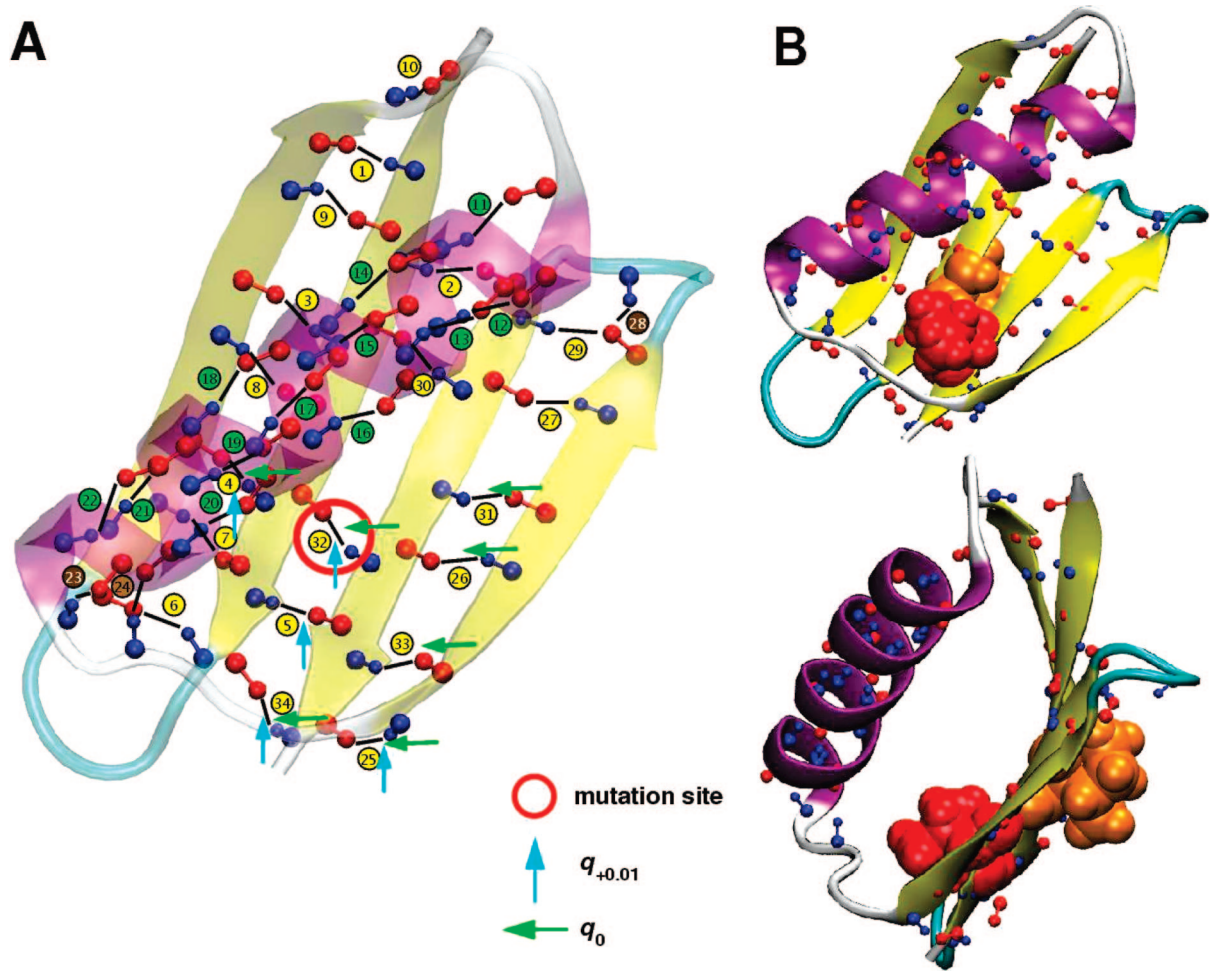



***Figure 7.*** Differences for calculated coupling constants for protein G mutants V54A and I6A. (A) Coupling constants which differ by more than 0.3 Hz compared to experimental values for the wild type protein indicated as arrows (green for $q_0$ and blue for $q_{+0.1}$). As can be seen, couplings away from J32 are also more or less affected depending on the charge set used. The mutation site is shown as a red circle. (B) Spatial orientation of the mutated sidechains V54 (red) and I6 (orange).

$^{3h}J_{NC'}$ coupling constants)[4] Hz for protein G, ubiquitin, and the Tudor domain, respectively. This is in almost quantitative agreement with the present work which yields rmsds of 0.14, 0.15, and 0.25 (0.21) Hz using the same force field and the same parametrization to calculate $^{3h}J_{NC'}$ coupling constants ($\beta_B = 3.20$, $q_0$). This is quite remarkable as the simulation approaches differ. Here, electrostatic cutoffs, multiple trajectories and extended simulation times (between 10 and 19 ns) are used compared to PME with 1 trajectory (except for protein G) and simulation times between 0.5 and 1.4 ns in the earlier study.[4]

Using biased MD simulations, it was previously shown that by including NMR energy constraints, (e.g., $^{3h}J_{NC'}$, NOE, and $S^2$) to the force field, shorter donor−acceptor distances and a more pronounced directionality of the H-bond are favored and lead to an ensemble that better reflects the experimentally measured one.[5] In this approach, one is interested in whether experimental constraints move the conformational ensemble from MD simulations toward the experimentally observed one. However, as the constraints are sequence- and amino acid-specific, little can be inferred about the particular merits and weaknesses of the underlying force field and predicting expected scalar couplings upon (point) mutation is not possible.

On the basis of the validation simulation discussed above, $^{3h}J_{NC'}$ coupling constants for particular protein G mutants were calculated. The donor and acceptor residues for H-bond number 32 (V54 → I6) were computationally mutated to alanine. For both systems independent 5 ns simulations were carried out with charge sets $q_0$ and $q_{+0.1}$. This choice was motivated by the observation that $q_0$ was previously found to be optimal for protein G and for all three proteins investigated here slightly increased charges ($q_{+0.1}$) improve the correlation between experiment and simulations. The estimated scalar couplings for mutant I6A are similar to the wildtype, whereas mutant V54A shows significant differences (see Figure 6). For the latter, with either charge set, couplings number J4, J5, J25, J26, J31, J32, J33, and J34 differ by more than 0.3 Hz from the experiment of the wild type protein (see Figure 7A). These couplings are located in the vicinity of the mutation site, more precisely in the $\beta$-sheet toward the sheet end. Using $q_0$, scalar couplings vanish for couplings number J25, J26, J32, J33, and J34 due to loss of $\beta$-sheet character in the affected strands. Such effects should be observable experimentally.

In conclusion, larger polarity (i.e., stronger donor−acceptor interactions) of the H-bond improves the calculated $^{3h}J_{NC'}$ coupling constants and the conformational ensemble sampled

from the MD simulations. The parameter $\beta_B$ is not universal across different proteins and different SSEs and optimized values moderately improve the calculated $^{3h}J_{NC'}$ coupling constants. The present work also suggests that scalar coupling constants can also be used to improve particular force field parameters. This will add a new aspect to force field development as NMR properties are intrinsically dynamical in nature. With additional measurements of scalar coupling constants in proteins (also on mutants of the ones discussed here), the available data could be considerably extended and a more exhaustive fit will be possible. Other dynamical quantities which are sometimes included are diffusion constants or, more recently, neutron structure factors[18] or NMR spin relaxation experiments.[19] In the future, it might be of interest to use force fields that take into account more details of the intermolecular interactions. One possibility is to use molecular mechanics with explicit proton transfer (MMPT) potentials which are 3-dimensional representations of model potentials for proton- or hydrogen-bonded motifs.[20,21] Another refinement that could be envisaged is to use higher multipole moments on the atoms involved in the H-bond. Such an extension has recently been found to improve the quality of simulations for infrared spectra of CO in myoglobin.[22]

**Acknowledgment.** We thank Prof. S. Grzesiek and Dr. I. Tubert-Brohman for valuable discussions and comments. This work was supported through SNF grant 200021-117810.

**Supporting Information Available:** Details on the structures for ab initio calculations, N···O′ distances, *B* factors, and rmsd derived from simulations and data on H-bonds studied in this work (PDF format). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Cordier, F.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 1601.

(2) Barfield, M. *J. Am. Chem. Soc.* **2002**, *124*, 4158.

(3) Markwick, P. R. L.; Sprangers, R.; Sattler, M. *J. Am. Chem. Soc.* **2003**, *125*, 644.

(4) Sass, H.-J.; Schmid, F.-F.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 5898–5903.

(5) Gsponer, J.; Hopearuoho, H.; Cavalli, A.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2006**, *128*, 15127–15135.

(6) 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, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J. E.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(7) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(8) Sprangers, R.; Groves, M. R.; Sinning, I.; Sattler, M. *J. Mol. Biol.* **2003**, *327*, 507.

(9) Selenko, P.; Sprangers, R.; Stier, G.; Buehler, D.; Fischer, U.; Sattler, M. *Nat. Struct. Biol.* **2001**, *8*, 27.

(10) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721.

(11) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906.

(12) Butterworth, S.; Lamzin, V. S.; Wigley, D. B.; Derrick, J. P.; Wilson, K. S. Anisotropic Refinement of a Protein G Domain at 1.1 Angstrom Resolution, in press.

(13) Cornilescu, G.; Ramirez, B. E.; Frank, M. K.; Clore, G. M.; Gronenborn, a. M.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 6275.

(14) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531.

(15) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836.

(16) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Chem. Phys.* **1977**, *23*, 327–341.

(17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2004.

(18) Thomas, J. L.; Tobias, D. J.; MacKerell, A. D., Jr. *J. Phys. Chem. B Lett.* **2007**, *111*, 12941.

(19) Showalter, S. A.; Brüschweiler, R. *J. Chem. Theor. Comp.* **2007**, *3*, 961–975.

(20) Lammers, S.; Meuwly, M. *J. Phys. Chem. A* **2007**, *111*, 1638.

(21) Lammers, S.; Lutz, S.; Meuwly, M. *J. Comput. Chem.* **2008**, *29*, 1048.

(22) Plattner, N.; Meuwly, M. *Biophys. J.* **2008**, *94*, 2505.

CT800241D

JCTC Journal of Chemical Theory and Computation

# Geometrical Preferences of the Hydrogen Bonds on Protein−Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations

Zhiguo Liu, Guitao Wang, Zhanting Li, and Renxiao Wang*

*State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Road, Shanghai 200032, P. R. China*

Received July 8, 2008

**Abstract:** We have conducted potential of mean force (PMF) analyses to derive the geometrical parameters of various types of hydrogen bonds on protein−ligand binding interface. Our PMF analyses are based on a set of 4535 high-quality protein−ligand complex structures, which are compiled through a systematic mining of the entire Protein Data Bank. Hydrogen bond donor and acceptor atoms are classified into several basic types. Both distance- and angle-dependent statistical potentials are derived for each donor−acceptor pair, from which distance and angle cutoffs are obtained in an objective, unambiguous manner. These donor−acceptor pairs are also studied by quantum mechanics (QM) calculations at the MP2/6−311++G** level on model molecules. Comparison of the outcomes of PMF analyses and QM calculations suggests that QM calculation may serve as an alternative approach for characterizing hydrogen bond geometry. Both of our PMF analyses and QM calculations indicate that C−H···O hydrogen bonds are relatively weak as compared to common hydrogen bonds formed between nitrogen and oxygen atoms. A survey on the protein−ligand complex structures in our data set has revealed that $C_\alpha$-H···O hydrogen bonds observed in protein−ligand binding are frequently accompanied by bifurcate N−H···O hydrogen bonds. Thus, the $C_\alpha$-H···O hydrogen bonds in such cases would better be interpreted as secondary interactions.

## 1. Introduction

Hydrogen bonding is probably the most important factor for maintaining the molecular structures and functions of various biological as well as chemical systems.[1,2] The very basic characteristics of a hydrogen bond is the D-H···A alignment, in which the hydrogen donor (*D*) is normally a strong electronegative atom such as nitrogen or oxygen, while the hydrogen acceptor (*A*) is another electronegative atom with at least one electron lone pair. Dissociation energy of a hydrogen bond may vary from 1 kcal/mol for a weak hydrogen bond such as C−H···O to 40 kcal/mol for a strong ionic hydrogen bond such as FH···$F^-$.[3] This feature endows hydrogen bonding an essential dual role: on one hand, hydrogen bonds are relatively weak compared to covalent bonds, thus they may form and break rapidly during the process of a conformational change or molecular recognition; on the other hand, due to the considerable strength and directional nature of hydrogen bonds, a desired specificity in structure can be eventually achieved.

An in-depth understanding of protein−ligand interactions has laid the foundation of structure-based drug design techniques, such as virtual screening,[4−6] *de novo* design,[7,8] and fragment-based design.[9−11] Hydrogen bonding is an essential factor in the binding process of a ligand molecule to its target protein. Many computational studies on this subject need to detect hydrogen bonds with rule-based algorithms, which rely on interpreting the relative positions and orientations, such as the D−A distance and the D-H-A angle, of two interacting chemical groups. Such algorithms are also implemented in some empirical scoring functions, such as the ones in LUDI,[12,13] FlexX,[14] ChemScore,[15,16] GlideScore,[17] SCORE,[18] and X-Score,[19] to estimate the contribution of hydrogen bonds to protein−ligand binding affinities. Char-

* Corresponding author phone: 86-21-54925128; e-mail: wangrx@mail.sioc.ac.cn.

10.1021/ct800267x CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/15/2008

acterization of hydrogen bonds is also an essential factor in some other theoretical studies, such as protein folding. Therefore, deduction of the preferred geometrical parameters of various types of hydrogen bonds is a meaningful goal for all these studies.

Geometrical parameters of hydrogen bonds can be derived from a statistical survey on a large number of crystal structures. Some studies of this kind have been reported before,[20−22] which were based on either the Cambridge Structure Database[23] (CSD) or the Protein Data Bank (PDB).[24] For the purpose of characterizing the hydrogen bonds on protein−ligand binding interface, apparently the latter approach is more straightforward. Due to the rapid progress in structural biology, the total number of available three-dimensional structures of biological macromolecules is growing constantly. While this manuscript is in preparation, over 50,000 structures have already been deposited in PDB. According to our previous analyses,[25,26] up to 40% of them can be classified as valid protein−ligand complexes. High-resolution structures of these protein−ligand complexes can serve as a solid basis for conducting statistical surveys regarding the hydrogen bonds on protein−ligand binding interface.

In this study, we have applied the potential of mean force (PMF) analysis on a large number of high-quality crystal structures of protein−ligand complexes to derive the geometrical preferences of various types of hydrogen bonds. It must be mentioned that the term "potential of mean force" could be confusing since it is actually more frequently used in other areas of molecular modeling, such as the molecular dynamics simulation of liquid phases. The PMF analysis applied in our study refers to the approach proposed by Sippl et al., which was originally applied to protein folding studies.[27,28] In recent years, this approach has been extended to the evaluation of protein−ligand binding by a number of scoring functions, such as PMF-Score,[29−31] DrugScore,[32,33] BLEEP,[34,35] SMoG,[36,37] DFIRE,[38,39] and M-Score.[40] The primary aim of our study is not to develop another PMF-based scoring function. Instead, we apply this approach to the characterization of the hydrogen bonds in protein−ligand binding, which is the first of this kind to the best of our knowledge. Our study covers common hydrogen bonds formed between oxygen and nitrogen atoms as well as C−H···O hydrogen bonds. To make comparison with the outcomes of our PMF analyses, we have also employed quantum mechanics (QM) calculations on some model molecules to characterize these hydrogen bonds. The geometrical parameters of various types of hydrogen bonds deduced in our study can be readily utilized by the empirical algorithms for perceiving hydrogen bonds in protein−ligand binding or protein folding studies.

## 2. Computational Details

**2.1. Preparation of Protein−Ligand Complex Structures.** Our statistical survey is conducted on a large set of high-quality structures of protein−ligand complexes. These complexes are selected throughout the entire Protein Data Bank (PDB) through a procedure similar to the one developed by us in the compilation of the PDBbind database.[25,26] This procedure can be described briefly as following. First, the composition of protein−ligand complex is considered. PDB entries which do not contain at least one protein molecule and one valid small-molecule ligand are filtered out. Here, a valid ligand must not be a cofactor/coenzyme (such as Heme, CoA, NAD, FAD, and their derivatives) or any component of an organic solvent and buffer. It also must not contain any uncommon elements, such as Be, B, Si, and metal atoms, and its molecular weight shall not exceed 1000. Note that oligopeptides (up to 9 residues) and oligonucleotides (up to 3 residues) are considered as valid small-molecule ligands in our study. Second, the quality of protein−ligand complex structure is considered. Only the protein−ligand complex structures which are determined through crystal diffraction with an overall resolution better than or equal to 2.5 Å are accepted. Finally, each qualified complex should be formed by one protein molecule with one ligand molecule in a binary manner, i.e. there should not be multiple ligands residing in close vicinity at the same binding site. In addition, covalently bound complexes are filtered out. All of the above examinations are conducted by a set of computer programs, which make judgments based on the contents of the original structural files downloaded from PDB. Manual inspections and adjustments are also employed whenever necessary.

We have screened the entire PDB (as released in January 2006) through the above procedure, and the outcome is a list of 4535 protein−ligand complexes. The structures of all of these complexes in the PDB format are downloaded from the RCSB PDB Web site (http://www.rcsb.org/pdb/). Each complex structure is then processed into appropriate formats for the convenience of subsequent analyses. In brief, each complex is split into a ligand molecule and a complete "biological unit" of the protein molecule, and they are saved in two separated files. Other components in the original PDB file, such as water and other solvent molecules, are ignored. The protein structure is sufficiently presented by the PDB format and thus does not need any additional treatment. The ligand structure, however, needs to be interpreted properly since the atom/bond type information is largely missing in the PDB format. The I-interpret program[41] is applied here to tackle this problem. This program interprets the chemical structure of a given organic molecule with a high accuracy merely based on the identities and coordinates of its component atoms. Each processed ligand is saved in the Mol2 format and is further manually inspected in the graphical interface of the Sybyl software[42] in order to detect any remaining problems in atom/bond types.

Since the primary aim of our study is to analyze hydrogen bonds, it is necessary to specify the explicit positions of hydrogen atoms on the protein and the ligand, which are normally absent in the original PDB structural files. In our study, the "standard" protonation states under neutral pH are applied to the ligand side, i.e. carboxylic, sulfonic, and phosphoric acid groups are set in deprotonated forms, while aliphatic amine groups, guanidine, and amidine groups are set in protonated forms. Hydrogen atoms are added onto the ligand accordingly with the Sybyl software. Situations on

***Table 1.*** Hydrogen Bond Donor and Acceptor Types Defined in Our PMF Analyses

| symbol | SMARTS string | description |
|---|---|---|
| | Donor Types | |
| OD.H | [$([#8]([#1])[#6])] | $sp^3$ oxygen atom in a hydroxyl group |
| ND.3 | [$([#7^3][#1])] | $sp^3$ nitrogen atom in an amine group, positively charged |
| ND.AM | [$([#7]([#1])[#6,#15,#16]=[#8]),$([#7]([#1])[#6]=[#16])] | nitrogen atom in an amide group |
| ND.PL3 | [$([#7;^2;D3][#1])] | $sp^3$ or $sp^2$ nitrogen atom with a triangle planar geometry[a] |
| CD.G | [$([#6][#1])] | generic carbon atom |
| CD.A | [$([#6]([#7])([#6]=[#8])[#1])] | alpha-carbon on an amino acid residue[b] |
| | Acceptor Types | |
| OA.2 | [$([#8]=*)] | $sp^2$ oxygen atom |
| OA.H | [$([#8;D2][#1])] | $sp^3$ oxygen atom in a hydroxyl group |
| OA.E | [$([#8;D2;H0])] | $sp^3$ oxygen atom in an ester or ether group |
| OA.NC | [$([#8;D1]~[#6,#15,#16]~[#8;D1])] | oxygen atom in a carboxylic group, negatively charged |
| NA.2 | [$([#7;D2]) | $sp^2$ nitrogen atom |

[a] Such as the nitrogen atom in pyrrole and the one in aniline. [b] Only applicable to protein molecules.

the protein side are more complicated since the protonation status of an amino acid residue may be affected by its surrounding environment. The PROPKA algorithm[43] is employed in our study to determine the protonation status of ionizable residues under neutral pH. This algorithm is chosen since its performance was the best in a recent benchmark.[44] Hydrogen atoms are then added onto the protein structure with the AMBER program[45] according to the predictions by PROPKA.

**2.2. Probing of Donor−Acceptor Pairs.** An in-house C++ program, PLHB, is developed based on the open source library in OpenBabel.[46] It is used to probe the donor−acceptor pairs on the binding interface of all of the protein−ligand complexes in our data set. Donor atoms and acceptor atoms are classified into several categories according to their chemical natures (Table 1). Combination of these donor and acceptor types covers most common hydrogen bonds observed between proteins and small-molecule ligands. The SMARTS chemical language[47] and the Programmable ATom TYper (PATTY) algorithm[48] are applied to this typing scheme. With SMARTS and PATTY, flexible and efficient atom type classification can be expressed in text strings that are interpretable to chemists.

Our PLHB program also computes the desired geometrical parameters of donor−acceptor pairs, including the D−A distance ($d$) and the D-H-A angle ($\theta$). Computation of the D-H-A angle needs the coordinates of hydrogen atoms, which are normally not available in the original structural files from PDB. Coordinates of the hydrogen atoms on most chemical groups can be reliably predicted with standard bond lengths, bond angles, and dihedral angles based on the hybridization state of their root atoms. An obvious exception is the hydrogen atom on a hydroxyl group (i.e., R-OH), which may have multiple possible positions around the R-O axis due to a low-energy rotation barrier. A simple searching algorithm is implemented in our PLHB program to tackle this problem: if a hydroxyl group is in close vicinity to an acceptor group on the counteracting molecule, the hydrogen atom on this hydroxyl group will be rotated around the R-O axis systematically to achieve the largest possible value of the D-H-A angle. The final coordinates of this hydrogen atom will be used in our statistical survey.

**2.3. Derivation of Statistical Potentials.** Pairwise potentials between donors and acceptors are derived from our data set of protein−ligand complexes through the potential of mean force (PMF) analysis. The basic idea beneath PMF analysis[27,28] is that statistically more populated configurations are energetically more favorable, and the ensemble of all accessible configurations are assumed to obey a Boltzmann distribution. In our study, the distance-dependent potential of each donor−acceptor pair is computed as

$$D_{ij}(d) = -RT\ln\left[\frac{m_0 + m_{ij}\dfrac{f_{ij}(d)}{g(d)}}{m_0 + m_{ij}}\right] \quad (1)$$

where $f_{ij}(d)$ is the relative probability of observing donor−acceptor pair $i$-$j$ at distance $d$, and $g(d)$ is the relative probability of observing a reference state at the same distance. Since our aim is to derive the geometrical preferences of hydrogen bonds over a nonspecific reference state, a reasonable choice of the reference state is all possible atom pairs, including van der Waals pairs as well as hydrogen bond pairs.

In eq 1, $f_{ij}(d)$ is computed as

$$f_{ij}(d) = \rho_{ij}(d)/\rho_{ij}(bulk) = \left(\frac{n_{ij}(d)}{4\pi d^2 \Delta d}\right) \Bigg/ \left(\frac{\sum_{D_{\min}}^{D_{\max}} n_{ij}(d)}{\int_{D_{\min}}^{D_{\max}} 4\pi d^2 \Delta d}\right) \quad (2)$$

And, $g(d)$ is computed as

$$g(d) = \rho_{all}(d)/\rho_{all}(bulk) = \left(\frac{n_{all}(d)}{4\pi d^2 \Delta d}\right) \Bigg/ \left(\frac{\sum_{D_{\min}}^{D_{\max}} n_{all}(d)}{\int_{D_{\min}}^{D_{\max}} 4\pi d^2 \Delta d}\right) \quad (3)$$

Here, $\rho_{ij}(d)$ is the numerical density of donor−acceptor pair $i$-$j$ observed at distance $d$, while $\rho_{ij}(bulk)$ is the numerical density of donor−acceptor pair $i$-$j$ observed throughout the entire sampling space. $\rho_{all}(d)$ and $\rho_{all}(bulk)$ are defined similarly, which are applied to all atom pairs. In our study, the lower bound ($D_{min}$) and the upper bound ($D_{max}$) of distance cutoff are set to 2.0 Å and 8.0 Å, respectively. In order to count the occurrence ($n_{ij}$) of donor−acceptor pair $i$-$j$ at a particular distance, the spherical sampling space centered at the donor atom is divided into multiple layers (Figure 1A). The bin width, *i.e.* $\Delta d$, is set to 0.1 Å. The

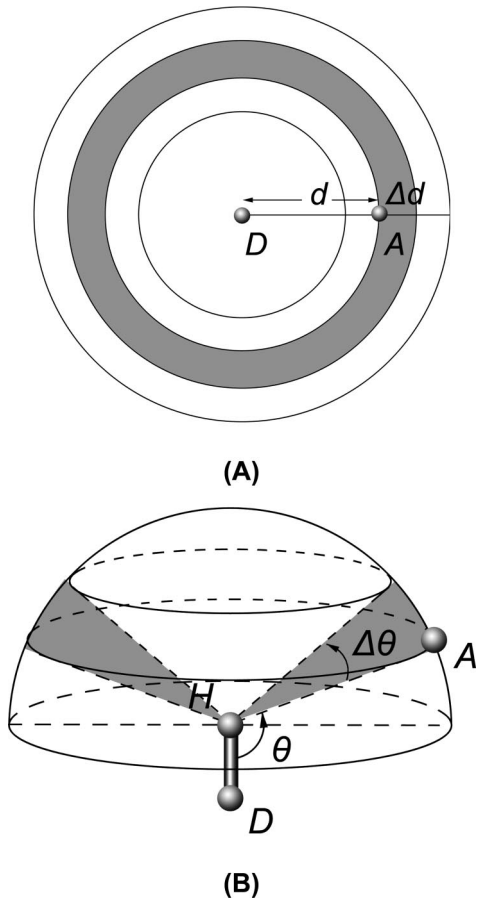



***Figure 1.*** Geometrical parameters considered in the deduction of (*A*) distance-dependent and (*B*) angle-dependent potentials. Volume of the shaded space is computed as (*A*) $\Delta V = 4\pi d^2_{\text{D-A}}\Delta d_{\text{D-A}}$ and (*B*) $\Delta V = {}^4/_3\pi d^3_{\text{H-A}}\sin({}^{\Delta\theta}/_2)\sin(\theta+{}^{\Delta\theta}/_2)$, respectively.

equation for computing the volume of each layer is given in the legend of Figure 1.

The introduction of $m_{ij}$ and $m_0$ in eq 1 is our extension to the standard algorithm for computing distance-dependent statistical potentials. $m_{ij}$ is the total occurrence of donor−acceptor pair *i-j* within a distance cutoff of 8.0 Å. It is typical that the occurrence of donor−acceptor pair *i-j* is really low at short distance, *i.e.* $f_{ij}(d)\rightarrow 0$ and $g(d)\rightarrow 0$. A small residual $m_0$ is introduced so that eq 1 will produce a meaningful value close to zero in such circumstances. The value of $m_0$ is set to 50 in our study, which is an arbitrary choice. In fact, no noticeable difference in the outcomes of eq 1 can be observed under different values of $m_0$ as long as $m_0$ is a relatively small number (see the Supporting Information, Part 2).

In our study, angle-dependent potentials of hydrogen bonds are also derived in a similar manner. The angle-dependent potential of donor−acceptor pair *i-j* is computed as

$$A_{ij}(\theta) = -RT\ln\left[\frac{m_0 + m_{ij}\dfrac{f_{ij}(\theta)}{g_{ij}(\theta)}}{m_0 + m_{ij}}\right] \tag{4}$$

where $f_{ij}(\theta)$ is the relative probability of observing donor−acceptor pair *i-j* at a particular angle $\theta$ when a hydrogen bond between them is possible, while $g_{ij}(\theta)$ is the relative probability of observing this donor−acceptor pair at the same angle regardless if a hydrogen bond between them is possible. Since $\theta$ is only relevant to donor−acceptor pairs, the reference state in eq 4 is different from the one in eq 1.

In eq 4, $f_{ij}(\theta)$ is computed as

$$f_{ij}(\theta) = \rho_{ij}(\theta)/\rho_{ij}(bulk) = \left(\frac{n_{ij}(\theta)_{d<D_{\max}}}{\frac{4}{3}\pi d^3\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)}\right)\Bigg/\left(\frac{\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}}{\frac{2}{3}\pi d^3}\right)$$
$$= \frac{n_{ij}(\theta)_{d<D_{\max}}}{2\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}} \tag{5}$$

Here $\rho_{ij}(\theta)$ is the numerical density of donor−acceptor pair *i-j* in hydrogen bonds observed at angle $\theta$. We use a distance cutoff ($D_{max}$) of 3.5 Å to decide if atoms *i* and *j* are close enough to form a hydrogen bond. This cutoff is chosen since it is approximately the sum of van der Waals radii of two heavy atoms in a common N−N, N−O, or O−O hydrogen bond. $\rho_{ij}(bulk)$ is the numerical density of donor−acceptor pair *i-j* in hydrogen bonds, i.e. when $d < D_{max}$, observed throughout the entire sampling space. The lower bound ($A_{min}$) and the upper bound ($A_{max}$) of angle $\theta$ are set to 90° and 180°, respectively. In order to count the occurrence of donor−acceptor pair *i-j* at a particular angle $\theta$, the semi-spherical sampling space centered at the hydrogen atom is divided into multiple cone-shaped sectors (Figure 1B). The bin width ($\Delta\theta$) is set to 5°. The equation for computing the volume of each sector is given in the legend of Figure 1.

In eq 4, $g_{ij}(\theta)$ is in fact computed using the same equation as $f_{ij}(\theta)$:

$$g_{ij}(\theta) = \rho_{ij}(\theta)/\rho_{ij}(bulk) = \frac{n_{ij}(\theta)_{d<D_{\max}}}{2\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}} \tag{6}$$

The only difference here is that the distance cutoff ($D_{max}$) is expanded to 8.0 Å. Thus, $g_{ij}(\theta)$ stands for the background probability of finding donor−acceptor pair *i-j* at angle $\theta$ regardless if they can form a valid hydrogen bond or not. $m_{ij}$ and $m_0$ in eq 4 have the same meanings as in eq 1.

**2.4. Quantum Mechanics Calculations.** We have also applied quantum mechanics calculations on model molecules to characterize hydrogen bonds. The model molecules used in our study are shown in Figure 2, which are selected to match the donor and acceptor types considered in our PMF analyses (Table 2). These model molecules are combined to produce a total of $4 \times 5 = 20$ donor−acceptor complexes. An initial configuration of each donor−acceptor complex is manually constructed first, in which the D−A distance (*d*) is set to 2.5 Å, the D-H-A angle ($\theta$) is set to 180°, and the lone pair on the acceptor atom is aligned with the A→D vector. An example is given in Figure 3A, showing how the initial configuration of the complex formed between a formylamide molecule (as the donor) and an acetone molecule (as the acceptor). The initial configuration is then subjected to structural optimization. Note that the association

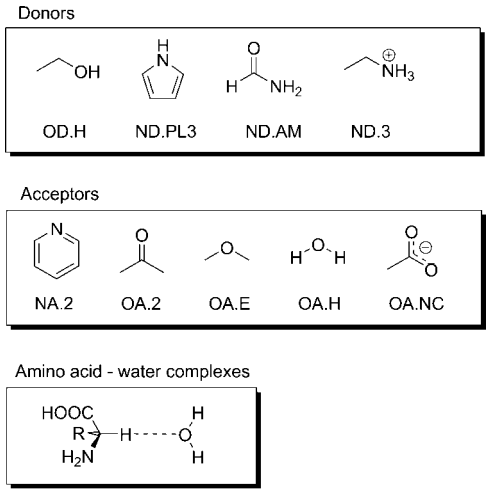



***Figure 2.*** Model molecules used in QM calculations. The symbol below each molecule is the corresponding donor or acceptor type that it represents.

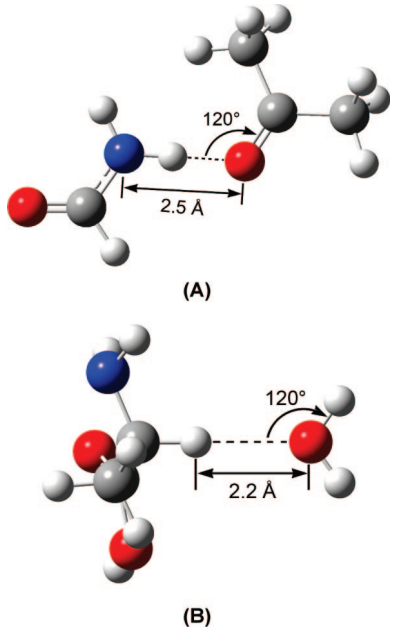



***Figure 3.*** Illustration of the initial configuration of a donor−acceptor complex subjected to structural optimization in our QM calculation. (A) The donor is the nitrogen atom on formylamide, while the acceptor is the oxygen atom on acetone. (B) The donor is the alpha-carbon atom on alanine, while the acceptor is the oxygen atom on a water molecule.

of two model molecules involves the desired hydrogen bond as well as other secondary interactions. In order to minimize the contribution of the latter so that the overall association energy is dominated by the desired hydrogen bond, angle $\theta$ is fixed as 180° during this process. All other degrees of freedom, including $d$, are fully relaxed.

The optimized complex structure is then subjected to frequency analysis. The association energy of the given complex is computed as

$$\Delta E_a^{298K} = E_{D-A}^{298K} - E_D^{298K} - E_A^{298K} + E_{BSSE} \quad (7)$$

Here, $\Delta E_{D-A}^{298K}$, $\Delta E_D^{298K}$, and $\Delta E_A^{298K}$ are the potential energies of the complex, the donor, and the acceptor at 298 K, respectively. They all include the contributions of zero point energies and thermal energies. $E_{BSSE}$ is the correction to the basis set superposition error (BSSE) computed with the counterpoise algorithm.[49] Based on the optimized complex structure, a potential energy scanning is also performed by varying $d$ systematically from 2.5 to 8.0 Å at an increment of 0.1 Å. All of the rotational degrees of freedom of two molecules are fixed so that the relative orientation of two molecules does not change during this process. The association energy of any particular configuration of the given complex during potential energy scanning is also computed with eq 7. The only difference is that zero point energy and thermal energy are not computed in each case because frequency analysis on every configuration is computationally too expensive.

We have also studied the hydrogen bonds involving the alpha-carbon atoms on amino acid residues through similar QM calculations. For this purpose, model molecules of 20 natural amino acids are constructed. Each model molecule is constructed as $NH_2CHRCOOH$, in which the amino group and the carboxylic group are set in neutral forms. To simulate the protonation states of amino acid residues on protein under neutral pH condition, the side chains of Asp and Glu are set in deprotonated forms, while the side chains of Lys and Arg are set in protonated forms. A water molecule is then used as the acceptor to probe the hydrogen bonding interaction with the $C_\alpha$ atom on each amino acid molecule. In the initial configuration of each amino acid−water complex, the distance between the oxygen atom on the water molecule and the hydrogen atom on the $C_\alpha$ atom is set to 2.2 Å. The $C_2$ axis of the water molecule is aligned with the $C_\alpha \rightarrow H$ vector. The H-$C_\alpha$-$C_\beta$-$X_\gamma$ dihedral angle of each amino acid molecule is set to 180° in order to avoid steric repulsions between the side chain and the water molecule (Figure 3B). This initial configuration is the subjected to structural optimization in which the $C_\alpha$-H-O angle is fixed as 180°. Computation of the association energy and the potential energy scanning for each amino acid−water complex are conducted through the same procedure described in the previous paragraph.

All calculations are performed using the GAUSSIAN 03 software[50] on an Intel Xeon 5345-based Linux cluster. Structural optimizations and single-point energy computations described above are all conducted at the MP2/6−311++G** level with frozen core approximation.

## 3. Results and Discussion

**3.1. Geometrical Preferences of Common Hydrogen Bonds Derived from PMF Analyses.** Among all of the geometrical parameters of a hydrogen bond, the D−A distance ($d$) and the D-H-A angle ($\theta$) are the most widely used. The preferred values of these geometrical parameters can be derived from statistical survey on a large number of crystal structures. For example, the D−A distance associated with the highest occurrence can be considered as the optimal distance for the hydrogen bond between D and A. This is in fact the standard approach adopted by some previous studies.[20−22] Our opinion is that PMF analysis is more

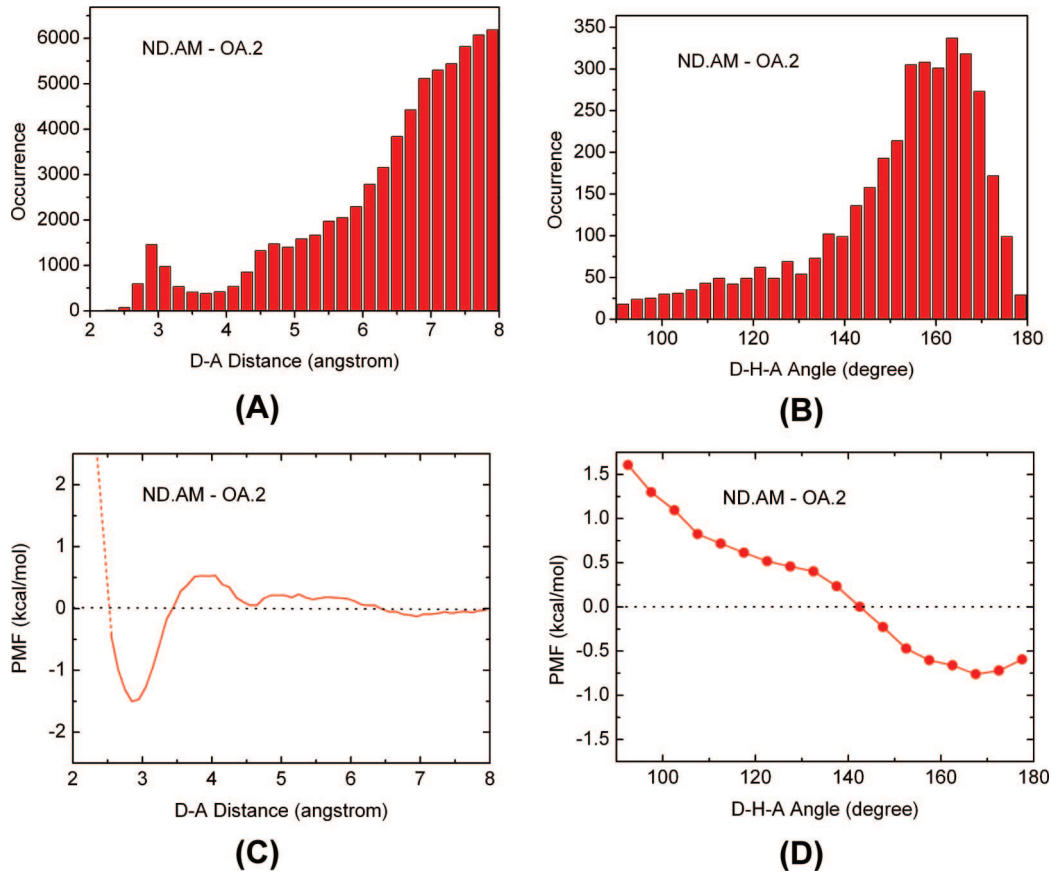



***Figure 4.*** Distributions of (*A*) the D−A distances and (*B*) the D-H-A angles of the ND.AM-OA.2 pair observed on our data set and the corresponding (*C*) distance-dependent and (*D*) angle-dependent PMF curves of this hydrogen bonding pair.

***Table 2.*** Optimal Interacting Distances and the Corresponding Statistical Potentials of Various Donor−Acceptor Pairs Derived from the Distance-Dependent PMF Analyses

| | acceptor | | | | |
|---|---|---|---|---|---|
| donor | NA.2 | OA.2 | OA.E | OA.H | OA.NC |
| OD.H | 2.7(−1.93)[a] | 2.7 (−1.44) | N/A[b] | 2.8(−1.84) | 2.7(−2.34) |
| ND.PL3 | N/A | 2.9 (−1.49) | 3.0(−0.75) | 2.9(−1.34) | 2.9(−1.89) |
| ND.AM | 3.1(−1.10) | 2.9(−1.50) | 3.0(−0.55) | 3.0(−1.05) | 2.8(−1.52) |
| ND.3 | N/A | 2.8(−1.61) | 3.1(−1.70) | 2.9(−2.03) | 2.9(−2.33) |
| CD.G | N/A | 3.3(−0.35) | N/A | 3.3(−0.50) | 3.3(−0.62) |
| CD.A | N/A | 3.3(−0.85) | N/A | 3.3(−0.24) | 3.3(−0.48) |

[a] The most preferred donor−acceptor interacting distance (in angstrom) of this atom pair; the number in brackets is the corresponding statistical potential (in kcal/mol) at this distance. [b] Reliable PMF curves cannot be obtained due to the low occurrence of this donor−acceptor pair in our data set.

appropriate for this purpose for two reasons. First, the optimal value of a certain geometrical parameter, e.g. the D−A distance, would be better located where the probability of finding this particular donor−acceptor pair reaches a maximum. The occurrence of this atom pair, however, does not necessarily reach its maximum at the same point. An appropriate correction is thus necessary since a larger D−A distance is associated with a larger sampling space ($\Delta V = 4\pi d^2 \Delta d$, see Figure 1A), and a larger sampling space is normally associated with higher occurrences. Second, the optimal value of a certain geometrical parameter would better be derived with consideration on its preference over an appropriate reference state. For example, when deriving the preference of the D−A distance of a hydrogen bond, it is appropriate to consider all atom pairs as the reference state (eq 1).

Here, we use a particular example, i.e. the hydrogen bonding pair ND.AM-OA.2, to further explain our approach and demonstrate its advantages. Distributions of the D−A distances and the D-H-A angles of this atom pair observed on our data set as well as the corresponding distance-dependent and angle-dependent statistical potentials derived through our approach are given in Figure 4. As one can see in Figure 4A, the occurrence of this atom pair has a local peak around 2.9 Å. This will be interpreted by a conventional counting-based approach as the optimal interacting distance ($d_0$) of this atom pair. Our distance-dependent PMF curve shows a sharp well at 2.9 Å, providing the same information. In addition, our PMF curve clearly shows a preferred interacting region for this atom pair, i.e. where $D_{ij}(d) < 0$ by eq 1. The upper bound of this region locates at 3.5 Å, where $D_{ij}(d) = 0$. It indicates that the hydrogen bond between

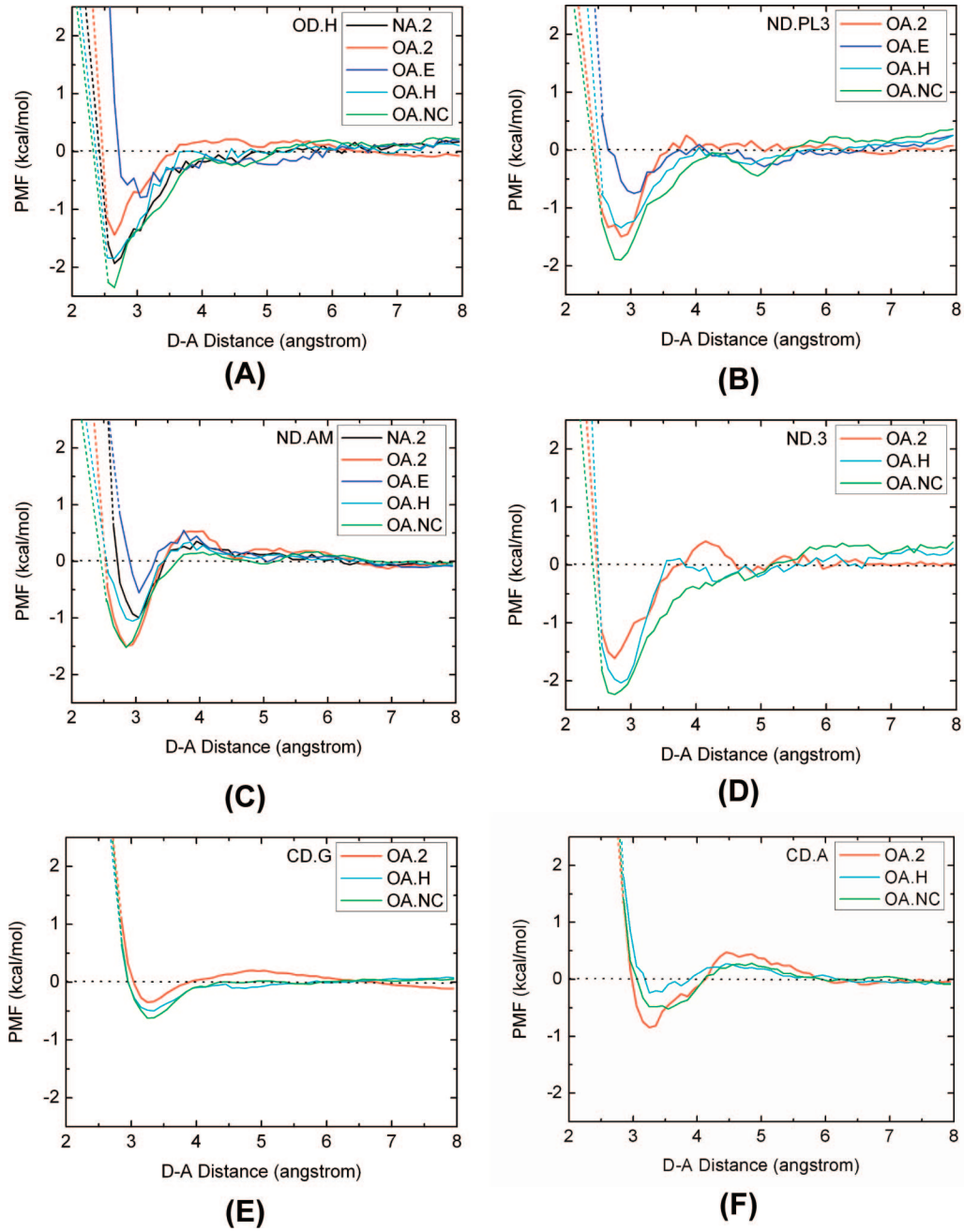



***Figure 5.*** Distance-dependent PMF curves of donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, (*E*) CD.G, and (*F*) CD.A. These curves at short distance, e.g. < 2.5 Å, are extrapolated due to the low occurrences of donor−acceptor pairs in this range. The extrapolated segments are rendered in dashed lines.

this atom pair becomes indistinguishable from the reference state at this particular distance. This critical distance, termed as $d^*$ in our study, can be interpreted as the distance cutoff of the given hydrogen bond. When the D−A distance is beyond this point, $D_{ij}(d)$ converges to the baseline quickly, whereas the occurrence of this atom pair keeps increasing with the D−A distance. Apparently, the conventional counting-based approach cannot deduce $d^*$ in an unambiguous manner.

The advantage of our approach is demonstrated even more clearly in the analysis of the D-H-A angle. As one can see in Figure 4B, the occurrence of the D-H-A angle at 180° is rather low, while the highest occurrence of this angle occurs around 165°. If relying on a simple count of occurrences, one may come to the conclusion that this kind of hydrogen bond prefers a somewhat twisted geometry rather than a perfect linear alignment. This type of statement is indeed witnessed in literature from time to time. However, our study points out that the low occurrence around 180° is simply due to a much smaller sampling space at this particular angle (Figure 1B). The angle-dependent PMF curve for this atom pair actually exhibits a relatively flat bottom between 160° and 180°, still supporting the linearity assumption. The critical D-H-A angle for this donor−acceptor pair to form a valid hydrogen bond, *i.e.* $\theta^*$, can be read from this curve as 140°, where $A_{ij}(\theta) = 0$ by eq 4. At this particular angle, whether this atom pair forms a hydrogen bond or not is indistinguishable even if they are close enough to be in a

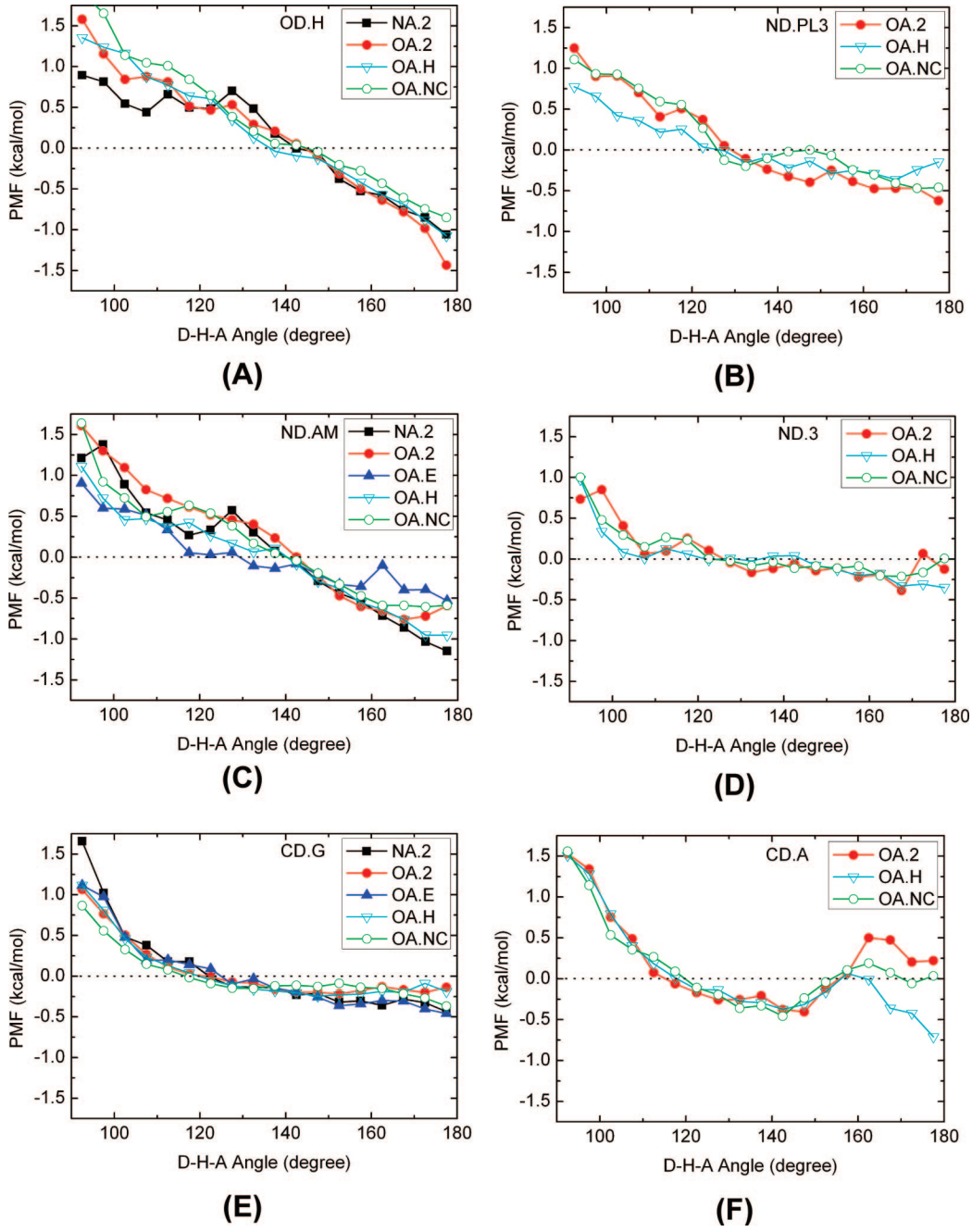



***Figure 6.*** Angle-dependent PMF curves of donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, (*E*) CD.G, and (*F*) CD.A.

hydrogen bonding range. This angle thus can be interpreted as the angular cutoff of the given type of hydrogen bond. Again, a conventional counting-based approach cannot deduce this parameter in an unambiguous manner.

The distance-dependent PMF curves of all of the donor−acceptor pairs considered in our study are shown in Figure 5. For common hydrogen bonds formed between oxygen and nitrogen atoms, a preferred interacting region is clearly shown on each curve. The most preferred interacting distances ($d_0$) of all donor−acceptor pairs are summarized in Table 2. As one can see in Figure 5 and Table 2, $d_0$ is somewhat different across various donor−acceptor pairs. As far as the hydrogen bonds containing the same type of acceptor are concerned, $d_0$ values in terms of four donor types are in an order of OD.H < ND.PL3 ≈ ND.AM ≈ ND.3. This is not surprising since an oxygen atom is more electronegative than a nitrogen atom, and therefore an oxygen atom as donor leads to a shorter hydrogen bond. No obvious trend is observed on the acceptor side though. As a rule of thumb, the average $d_0$ values of an O−O, O−N, and N−N hydrogen bond are 2.7 Å, 2.9 Å, and 3.1 Å, respectively. Note that $d_0$ values of the donor−acceptor pairs containing OA.E tend to be larger by 0.1−0.2 Å than those containing other types of acceptor atoms (OA.2 or OA.H). By our definition, OA.2 or OA.H is covalently connected with only one heavy atom, while OA.E is covalently connected with two. The steric repulsion introduced by the neighboring atoms of OA.E could account for the slightly longer hydrogen bonds. As for critical distances, our results show that $d^*$ for an O−O, O−N, and N−N hydrogen bond typically ranges

from 3.5 to 4.0 Å (Figure 5). The $d^*$ for the ND.3-OA.NC pair is significantly longer (~5.0 Å), indicating that the interaction between these two particular atom types is electrostatic rather than a typical hydrogen bond.

Another notable feature of the distance-dependent PMF curves shown in Figure 5 is that they are not monotonous when $d$ is larger than the optimal interacting distance ($d_0$). Instead, some fluctuations around the baseline are observed typically where $d > d^*$. Such fluctuations are the consequence of packing effects. Atoms cannot distribute freely in the three-dimensional space because they are all restricted by some chemical bonds. Moreover, each atom has a certain size so that even two unbound atoms need to be apart by a certain distance. In other words, atoms are packed in discrete layers rather than a perfect continuous manner. The same phenomenon is also seen in virtually every set of statistical potentials for protein−ligand binding derived by other researchers.[29−40]

Angular preferences are also very important for hydrogen bonds because they are directional in nature. Our definition of the angle-dependent potentials is an extension to the standard PMF approach, which is normally applied to the derivation of distance-dependent potentials. The angle-dependent PMF curves of various types of hydrogen bonds derived in our study are given in Figure 6. Our results show that all of the hydrogen bonds formed between oxygen and nitrogen atoms have a clear preference toward a linear alignment of D-H-A, *i.e.* $\theta = 180°$. Interestingly, a previous study by Desiraju et al. on a data set consisting of 28 protein−ligand complexes shows that hydrogen bonds formed between protein and ligand exhibited certain deviations from linearity.[51] In their study, a cone-correction[52] was applied to describe angular preference, an approach similar to ours. The discrepancy between our observation and theirs is probably due to the small data set employed in their study. As for the critical angle, we have observed that this parameter is basically determined by the donor type (Figure 6). As a rule of thumb, the D-H-A angular cutoffs for OD.H, ND.3, ND.PL3, and ND.AM are 140°, 125°, 125°, and 140°, respectively. These angle parameters, together with the distance parameters discussed above, can be readily utilized by empirical algorithms for perceiving hydrogen bonds.

**3.2. Comparison of the Outcomes of QM Calculations and PMF Analyses.** In the past two decades or so, a number of studies have employed QM calculations to characterize hydrogen bonds for various purposes.[53−59] In this study, we have followed this approach to explore the geometrical preferences of various types of hydrogen bonds. It needs to be emphasized that our QM calculations on simple model systems are independent from our PMF analyses on a large number of protein−ligand complex structures. Our purpose is to investigate if these two different approaches can achieve any consensus in terms of the geometrical and energetic properties of hydrogen bonds. One can also get a better understanding of both the strengths and shortcomings of these two approaches through this comparison.

The computed association energy of each donor−acceptor complex as a function of the D−A distance, i.e. the outcomes of potential energy scanning, is plotted in Figure 7. One can see that these distance-dependent energy curves and the corresponding distance-dependent PMF curves resemble in a qualitative manner: both types of curves exhibit a maximal interaction at a certain distance, and they converge to zero at a large distance. The optimal donor−acceptor interacting distances ($d_0$) of all donor−acceptor complexes are summarized in Table 3. For the neutral hydrogen bonds containing the same type of acceptors, $d_0$ values are in a clear order of OD.H < ND.PL3 < ND.AM, although the difference is subtle, while for the neutral hydrogen bonds containing the same type of donors, $d_0$ values are essentially the same across various acceptor types. As for "charged" hydrogen bonds, i.e. those containing ND.3 or OA.NC, $d_0$ values are generally shorter by 0.2−0.3 Å than those of neutral hydrogen bonds. Notably, the absolute values of $d_0$ of various hydrogen bonds given by our QM calculations agree well with those reported by Marian et al. in a recent study.[59]

Comparing the $d_0$ values given by PMF analyses (Table 2) with those given by QM calculations (Table 3), one can see that they match well in most cases, especially the ones associated with the same type of nitrogen donors. The discrepancy between two sets of data ranges typically between 0 and 0.3 Å (<0.1 Å for eight hydrogen bonds; 0.1−0.2 Å for four; >0.2 Å for four). Considering the limited resolution of the crystal structures in our data set, an overall agreement at this level is acceptable. Nevertheless, we have also noticed that the $d_0$ values of some neutral hydrogen bonds given by PMF analyses tend to be shorter by 0.1−0.3 Å than those given by QM calculations. This can be ascribed to the frequent occurrence of bifurcated hydrogen bonds in protein−ligand binding. Bifurcated hydrogen bonds are stronger intermolecular interactions and therefore are associated with shorter donor−acceptor distances as compared to single hydrogen bonds. In contrast, the model systems considered in our QM calculations allow only one hydrogen bond in each donor−acceptor complex. Thus, the results of both PMF analyses and QM calculations should be interpreted in their own contexts.

The hydrogen bonding energy is a more subtle issue. The association energies ($\Delta E^{298K}_{a}$) of all donor−acceptor complexes given by our QM calculations are also summarized in Table 3. The energies of the hydrogen bonds formed between neutral oxygen and nitrogen atoms range from −2.6 to −6.2 kcal/mol. The energies of the hydrogen bonds containing charged atoms (ND.3 and OA.NC) range from −15.6 to −25.3 kcal/mol, roughly 4−5 times more negative than those of neutral hydrogen bonds. This is understandable since electrostatic interactions are much more significant in the cases of charged hydrogen bonds. In contrast, the interaction potentials read from the distance-dependent PMF curves of various donor−acceptor pairs scatter in a relatively narrow range, i.e. −0.5 to −2.5 kcal/mol (Table 2). No obvious correlation can be found between the energy data in Tables 2 and 3. Particularly, the interaction potentials of charged hydrogen bonds are not significantly more negative than those of neutral hydrogen bonds.

One may argue that unlike QM computed energies, statistical potentials produced by an equation like eq 1 include solvation effects implicitly since they are derived

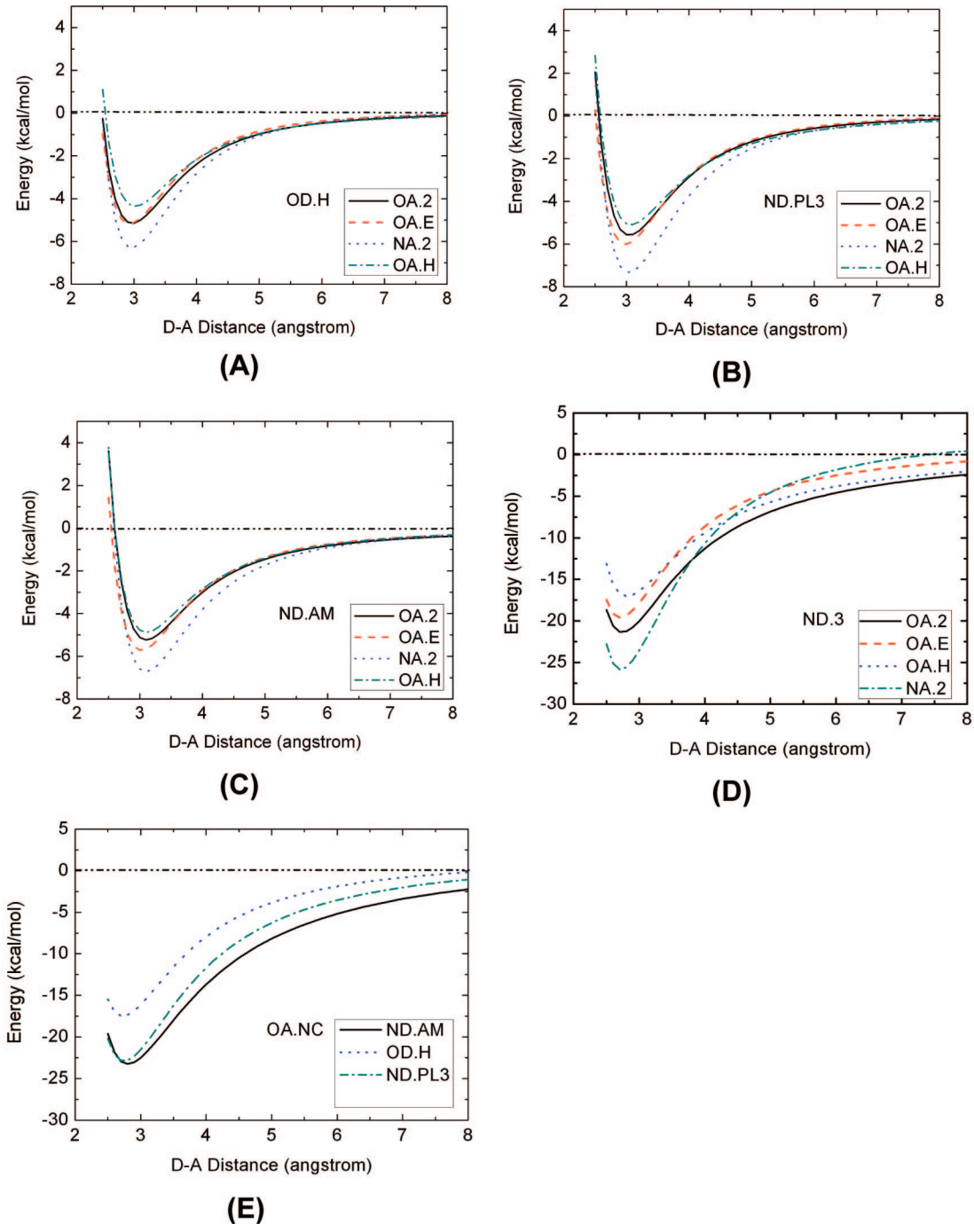



***Figure 7.*** Association energies of the donor−acceptor complexes containing donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, and acceptor type (*E*) OA.NC calculated at the MP2/6−311++G** level.

***Table 3.*** Optimal Interacting Distances and the Corresponding Association Energies of Various Donor−Acceptor Complexes Calculated at the MP2/6−311++G** Level

| | acceptor[a] | | | | |
|---|---|---|---|---|---|
| donor[a] | NA.2 | OA.2 | OA.E | OA.H | OA.NC |
| OD.H | 2.97(−4.62)[b] | 2.97(−3.69) | 2.93(−2.93) | 3.02(−2.69) | 2.73(−16.02) |
| ND.PL3 | 3.04(−6.17) | 3.04(−3.40) | 2.97(−4.97) | 3.07(−4.13) | 2.74(−22.58) |
| ND.AM | 3.10(−4.45) | 3.10(−3.02) | 3.02(−4.07) | 3.10(−3.99) | 2.80(−21.72) |
| ND.3 | 2.72(−25.32) | 2.74(−20.64) | 2.71(−18.95) | 2.81(−15.61) | N/A[c] |

[a] The corresponding model molecules are shown in Figure 2. [b] The optimal interacting distance (in angstrom) between the donor atom and the acceptor atom; the number in brackets is the association energy of this complex at 298 K (in kcal/mol), including zero point energy and thermal energy corrections. [c] In this case, one hydrogen atom translocates from the positively charged nitrogen atom to the negatively charged oxygen atom after energy minimization of the initial configuration. The computed energy does not reflect the formation of a hydrogen bond and thus is not reported here.

from protein−ligand complex structures that are fully solvated. Formation of a charged hydrogen bond is accompanied with larger desolvation penalties, and thus the net gain is not more significant than the one of a neutral hydrogen bond. Our opinion is that this statement may not be true. Technically, eq 1 only gives the relative probability of finding a given atom pair at any particular distance. The physical basis of using the logarithm of such probability as interaction potential is actually vague. For example, one can see in Table 2 that the statistical potentials of the hydrogen bonds containing OA.E ($sp^3$ oxygen atom in an ether or ester group) as acceptor are consistently less negative than those containing OA.H ($sp^3$ oxygen atom in a hydroxyl group) as acceptor (−0.75 vs −1.34, −0.55 vs −1.05, and −1.70 vs −2.03). However, our QM calculations (Table 3) indicate that interaction energies of the hydrogen bonds of the former type are comparable or even more negative than those of the latter type in vacuum (−4.97 vs −4.13, −4.07 vs −3.99, and −18.95 vs −15.61). Considering that the desolvation energies of dimethyl ether ($CH_3OCH_3$) and methanol ($CH_3OH$) are 1.92 and 5.11 kcal/mol, respectively, it is not reasonable to expect that the net energy of a hydrogen bond of the former type is less negative than the counterpart of the latter type. The hydrogen bonds of the former type are associated with less negative statistical potentials simply because their occurrence is considerably lower than the one of the latter type in our data set (see the Supporting Information).

In our study, we have investigated hydrogen bonds through two different approaches. In contrast to QM calculation on simple model molecules, PMF analysis is capable of characterizing the hydrogen bonds formed during protein−ligand binding *in situ*. This approach can be used to deduce some geometrical parameters that are useful for scoring function or force field development, such as the distance and angle cutoffs of various hydrogen bonds. Such parameters are difficult to obtain through QM calculations. Nevertheless, the PMF analysis approach also has its shortcomings. One major problem is that reliable PMF potentials cannot be obtained for the atom pairs with low occurrences, e.g. those labeled as "N/A" in Table 2. Another problem is that the outcomes of PMF analysis are statistical averages, which can be ambiguous sometimes. In contrast, the outcomes of QM calculation are usually straightforward to interpret since they are obtained on idealized model systems. Also, QM calculation is technically applicable to any appropriate model systems. However, both the geometries and energies obtained through QM calculation on model molecules in vacuum need to be validated with caution in the context of protein−ligand binding. Therefore, our opinion is that PMF analysis and QM calculation are two complementary approaches to characterize the geometries of the hydrogen bonds formed in protein−ligand binding. It is however not appropriate to compare statistical potentials with QM-calculated energies since they are derived from different contexts.

**3.3. On C−H···O Hydrogen Bonds.** Some previous studies have reported that uncommon hydrogen bonds, especially C−H···O, are observed in the crystal structures of small organic molecules[60,61] and proteins[62−64] as well as protein−ligand and protein−protein complexes.[65−67] Some researchers state that the frequent occurrence of C−H···O hydrogen bonds indicates their essential role in the stability of protein structures.[68,69] A number of in-depth QM studies on C−H···O═C interactions have been done by Dixon et al.,[70−73] who demonstrate that such interactions can be fairly strong. However, some other studies provide conflicting conclusions that C−H···O hydrogen bonds may not make significant contributions to the stability of protein structures.[74,75] Conflicting viewpoints on the role of C−H···O hydrogen bonds in protein−ligand binding can also be found in the literature.[65,76,77] In this study, we have performed both PMF analyses and QM calculations on the C−H···O hydrogen bonds in order to investigate their elusive role.

As one can see from Figure 5E,F, the distance-dependent PMF curves of carbon donors, i.e. CD.G and CD.A, also exhibit a preferred interacting region, a feature similar to those of nitrogen and oxygen donors. However, the potential wells observed on these PMF curves are generally shallower than those of common hydrogen bonds. The most preferred interacting distance ($d_0$) involving carbon donors is considerably larger (∼3.3 Å) (Table 2), which agrees well with a previous survey of hydrogen bonds in protein−protein interaction.[67] Note that this distance is actually close to the sum of van der Waals radii of carbon and oxygen. The angle preferences of C−H···O hydrogen bonds are also different from those of common hydrogen bonds. The angle-dependent PMF curve of donor type CD.G is somewhat flat between 140° and 180° (Figure 6E). No strong preference toward a linear alignment of C−H···O is observed. All of these observations suggest that the C−H···O bonds on protein−ligand binding interface are quite different from common hydrogen bonds. They resemble nonspecific van der Waals contacts more closely from a statistical point of view.

We have paid special attention to the alpha-carbons on amino acid residues since they are more acidic than general carbon atoms and thus are more likely to form genuine C−H···O hydrogen bonds. Unexpectedly, the $C_\alpha$-H-O angle shows a clear preference to 140° as indicated by the angle-dependent PMF curve of donor type CD.A (Figure 6F). We thus suspect that in such a case, the acceptor atom on the ligand side may form a bifurcated hydrogen bond with a nearby amide group on the protein backbone, which consequently forces the $C_\alpha$-H-O angle to deviate from linearity. In order to prove this, we have re-examined the entire data set for all of the amino acid residues involved in $C_\alpha$-H···O hydrogen bonds with the ligand side. Distribution of the H-$C_{i,\alpha}$-$C_i$-$N_{i+1}$ dihedral angles of these residues reveals that the most populated value of this angle is around 30° (Figure 8A), an angle quite suitable for the proximity of the hydrogen atoms on $C_{i,\alpha}$ and $N_{i+1}$. Another piece of supportive evidence comes from the Ramachandran plot of these residues (Figure 8B). One can see that most of these residues reside in $\beta$-strands, which have appropriate $\Psi$ dihedral angles facilitating the access to the hydrogen atoms on $C_{i,\alpha}$ and $N_{i+1}$ by the same acceptor atom on the ligand side. As a matter of fact, among all the $C_{i,\alpha}$-H···O bonds observed in our data set, over 73% of them (989 in 1351) are found to be accompanied with a bifurcated $N_{i+1}$-H···O hydrogen bond. Considering that C−H···O hydrogen bonds are generally

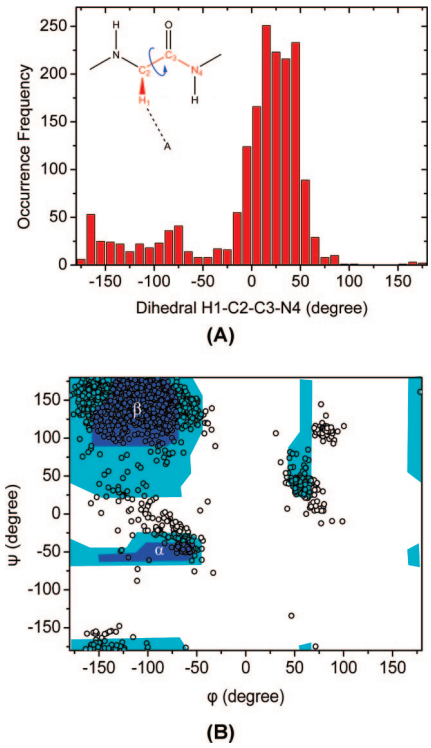



***Figure 8.*** (*A*) Distribution of the $H_1$-$C_2$-$C_3$-$N_4$ dihedral angles of all amino acid residues which are involved in $C_\alpha$-H···O hydrogen bonds. (*B*) Ramachandran plot of these residues. In both cases, a $C_\alpha$-H···O hydrogen bond will be counted if the $C_\alpha$-O distance is shorter than 3.6 Å and the $C_\alpha$-H-O angle is larger than 90°. Glycines are excluded in this survey since they have no side chains.

much weaker than regular hydrogen bonds, it is appropriate to interpret the $C_\alpha$-H···O hydrogen bond as a resultant phenomenon in such a case. It should be emphasized that our analysis does not necessarily rule out the contribution of all C−H···O bonds. For example, when a binding pocket is predominantly hydrophobic, even one C−H···O hydrogen bond can be critical for achieving the specific binding of a ligand.[78] Our analysis prompts that one needs to be extremely careful when interpreting the role of a C−H···O bond. For example, if ignoring the frequent co-occurrence of $C_{i,\alpha}$-H···O and $N_{i+1}$-H···O hydrogen bonds, one may come down to the wrong conclusion that a $C_\alpha$-H···O hydrogen bond is an independent factor common in protein−ligand binding.

In our QM study of $C_\alpha$-H···O hydrogen bonds, a water molecule is used as a probe to interact with the $C_\alpha$ atoms on all 20 types of amino acids. The optimal interacting distance and the corresponding association energy of each amino acid−water complex are summarized in Table 4. The optimal $C_\alpha$···O interacting distances typically range from 3.5 to 3.6 Å for most amino acids. Notably, these values are again greater by ∼0.2 Å than the counterparts derived from our PMF analyses. This discrepancy is understandable since there is no bifurcate $N_{i+1}$-H···O hydrogen bond to bring the water molecule closer to the $C_\alpha$ atom in the model systems considered in our QM calculations, whereas it is a frequently occurring event on real protein−ligand complexes.

***Table 4.*** Properties of the $C_\alpha$-H···O Hydrogen Bonds between 20 Amino Acids and Water Molecules Calculated at the MP2/6−311++G** Level

| model | $d_{O\cdots H}$ (Å)[a] | $d_{O\cdots C}$ (Å)[b] | $\triangle d_{C-H}$ (Å)[c] | $\triangle E_a^{298K}$ (kcal/mol)[d] | $\triangle E_a^{298K}$ (kcal/mol)[e] |
|---|---|---|---|---|---|
| GLY | 2.50 | 3.60 | −0.0023 | −1.70 | −1.00 |
| ALA | 2.45 | 3.55 | −0.0028 | −1.53 | −1.41 |
| VAL | 2.49 | 3.58 | −0.0033 | −1.81 | −1.13 |
| LEU | 2.49 | 3.59 | −0.0029 | −1.65 | −1.55 |
| ILE | 2.44 | 3.53 | −0.0016 | −1.86 | −1.79 |
| PHE[f] | 2.43 | 3.52 | −0.0014 | −1.78 | N/A |
| TYR[f] | 2.43 | 3.52 | −0.0014 | −1.75 | N/A |
| TRP[f] | 2.44 | 3.54 | −0.0013 | −1.66 | N/A |
| CYS | 2.45 | 3.54 | −0.0030 | −2.16 | −2.03 |
| MET | 2.46 | 3.56 | −0.0032 | −1.91 | −1.81 |
| ASN | 2.43 | 3.53 | −0.0025 | −2.41 | −1.68 |
| GLN | 2.47 | 3.57 | −0.0016 | −1.53 | −0.77 |
| SER | 2.46 | 3.55 | −0.0010 | −1.82 | −0.52 |
| THR | 2.46 | 3.56 | −0.0017 | −2.19 | −1.45 |
| HIS | 2.37 | 3.46 | −0.0021 | −2.64 | −1.91 |
| ARG[f] | 2.36 | 3.46 | −0.0024 | −4.27 | N/A |
| LYS | 2.37 | 3.46 | −0.0028 | −4.14 | −2.79 |
| ASP | 2.94 | 4.05 | 0.0002 | −3.39 | −3.37 |
| GLU | 3.04 | 4.14 | 0.0004 | −2.74 | −2.74 |

[a] Distance between the oxygen atom on water and the hydrogen atom on the $C_\alpha$ atom. [b] Distance between the oxygen atom on water and the $C_\alpha$ atom. [c] Change in the $C_\alpha$-H bond length upon the formation of the $C_\alpha$-H···O hydrogen bond. [d] Association energy of the amino acid−water complex at 0 K. [e] Association energy of the amino acid−water complex at 298 K, including zero point energy and thermal energy corrections. [f] Complete convergence is not achieved in structural optimization due to the complexity of the model system. Consequently, zero point energy and thermal energy corrections are not computed since the frequency analysis is not feasible in this case.

The association energies of amino acid−water complexes produced by our QM calculations vary significantly among different types of amino acids (Table 4). As for the complexes involving neutral amino acids, the $\Delta E^{298K}{}_a$ values range between −0.5 and −2.0 kcal/mol. In this regard, these $C_\alpha$-H···O hydrogen bonds are generally weaker as compared to the common hydrogen bonds formed by nitrogen and oxygen atoms (Table 3). In contrast, the complexes involving charged amino acids, including Arg, Lys, Asp, and Glu, have considerably more negative $\Delta E^{298K}{}_a$ values between −2.7 and −3.4 kcal/mol, close to the level of common hydrogen bonds. These enhanced association energies should be attributed to the charge-dipole interactions between these amino acids and water molecules. The association energy is plotted in Figure 9 as a function of $C_\alpha$···O distance for three selected amino acids, i.e. Ala, Thr, and Lys. One can see that as the $C_\alpha$···O distance increases, the association energy of the Lys-water complex converges to the baseline much slower than that of the Ala-water or Thr-water complex. This is a typical characteristic of long-range electrostatic interactions. Moreover, in the cases of negatively charged amino acids, i.e. Asp and Glu, the water molecule actually turns over after structural optimization, pointing its electron lone pairs rather than hydrogen atoms to the amino acid. The hydrogen atoms represent the positive end of the water dipole, and thus turning over of the water molecule will facilitate its charge-dipole interaction with the amino acid. Turning over of the water molecule, of course, eliminates the possibility of forming the $C_\alpha$-H···O bond. Based on these

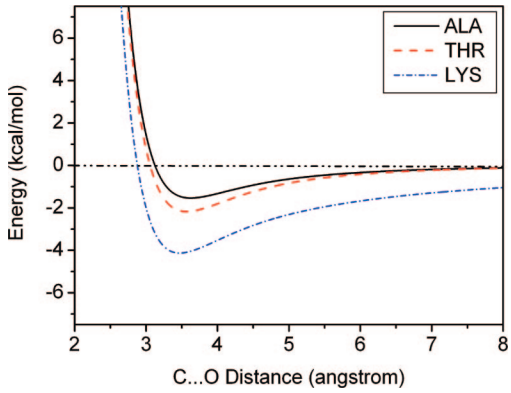



***Figure 9.*** Association energies of three selected amino acid−water complexes calculated at the MP2/6−311++G** level.

observations, we conclude that the interactions between charged amino acids and water molecules are dominated by electrostatic interactions rather than the $C_\alpha$-H···O hydrogen bond. The contribution of a C−H···O hydrogen bond in these cases may not be as significant as what some previous studies have suggested.[79]

## 4. Conclusions

We have analyzed the geometrical preferences of various types of hydrogen bonds found on protein−ligand binding interface. Our PMF analyses are based on a large set of high-quality protein−ligand complex structures, which is compiled through a systematic mining of the entire PDB. We have demonstrated that one can obtain both distance- and angle-dependent statistical potentials for a given type of hydrogen bond, from which distance and angle cutoffs can be obtained in an objective, unambiguous manner. Such geometrical parameters can be readily utilized by empirical algorithms for perceiving hydrogen bonds. The results given by our PMF analyses are also compared with those given by QM calculations on model molecules. The optimal interacting distances given by the two approaches are basically in accordance with each other except for a few cases. This suggests that QM calculation may serve as an alternative approach for characterizing hydrogen bond geometry especially when PMF analysis is not applicable. Nevertheless, no obvious correlation has been observed between the statistical potentials given by PMF analyses and the association energies given by QM calculations. It is not appropriate to validate QM energies with statistical potentials and *vice versa*. Both of our PMF analyses and QM calculations indicate that C−H···O hydrogen bonds are relatively weak as compared to common hydrogen bonds formed between nitrogen and oxygen atoms. In particular, our survey on protein−ligand complex structures reveals that the relatively frequent occurrence of $C_\alpha$-H···O hydrogen bonds in protein−ligand binding is largely due to the coexistence of bifurcate N−H···O hydrogen bonds. Thus, the $C_\alpha$-H···O hydrogen bonds in such cases would be better interpreted as secondary interactions.

**Acknowledgment.** The authors are grateful for the financial support from the Chinese National Natural Science Foundation (Grants 20502031 & 20772149), the Chinese Ministry of Science and Technology (the 863 high-tech project, Grant 2006AA02Z337), and the Science and Technology Commission of Shanghai Municipality (Grants 06PJ14115 & 074319113).

**Supporting Information Available:** Some raw data of the PMF analyses and QM calculations conducted in this study as well as more descriptions and discussion. This material is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Wormer, P. E.; van Der Avoird, A. *Chem. Rev.* **2000**, *100*, 4109–4144.

(2) Hobza, P.; Havlas, Z. *Chem. Rev.* **2000**, *100*, 4253–4264.

(3) Steiner, T. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 49–76.

(4) Ghosh, S.; Nie, A.; An, J.; Huang, Z. *Curr. Opin. Chem. Biol.* **2006**, *10*, 194–202.

(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(6) Jain, A. N. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 396–403.

(7) Schneider, G.; Fechner, U. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.

(8) Dean, P. M.; Lloyd, D. G.; Todorov, N. P. *Curr. Opin. Drug Discovery. Dev.* **2004**, *7*, 347–353.

(9) Hajduk, P. J.; Greer, J. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.

(10) Ciulli, A.; Abell, C. *Curr. Opin. Biotechnol.* **2007**, *18*, 489–496.

(11) Verdonk, M. L.; Hartshorn, M. J. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 404–410.

(12) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

(13) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.

(14) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470–489.

(15) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(16) Murray, C. W.; Auton, T. R.; Eldridge, M. D. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503–519.

(17) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(18) Pei, J.; Wang, Q.; Zhou, J.; Lai, L. *Proteins* **2004**, *57*, 651–664.

(19) Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11–26.

(20) Taylor, R.; Kennard, O. *Acc. Chem. Res.* **1984**, *17*, 320–326.

(21) Mills, J. E. J.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607–622.

(22) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.

(23) Allen, F. H. *Acta. Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(25) Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(26) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(27) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.

(28) Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229–235.

(29) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791–804.

(30) Muegge, I. *J. Comput. Chem.* **2001**, *22*, 418–425.

(31) Muegge, I. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(32) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.

(33) Velec, H. F.; Gohlke, H.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6296–6303.

(34) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. *J. Comput. Chem.* **1999**, *20*, 1165–1176.

(35) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. *J. Comput. Chem.* **1999**, *20*, 1177–1185.

(36) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770–2780.

(37) Dominy, B. N.; Shakhnovich, E. I. *J. Med. Chem.* **2004**, *47*, 4538–4558.

(38) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005**, *48*, 2325–2335.

(39) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. *Proteins* **2004**, *56*, 93–101.

(40) Yang, C. Y.; Wang, R.; Wang, S. *J. Med. Chem.* **2006**, *49*, 5903–5911.

(41) Zhao, Y.; Cheng, T.; Wang, R. *J. Chem. Inf. Model* **2007**, *47*, 1379–1385.

(42) *Sybyl, version 7.3*; Tripos, Inc.: St. Louis, MO 63144, U.S.A., 2006.

(43) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(44) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. *BMC Biochem*. [Online] **2006**, *7*, Article 18. http://www.biomedcentral.com/1471−2091/7/18/ (accessed Nov 27, 2007).

(45) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.;Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.;Kollman, P. A. *AMBER, version 9.0*; University of California, San Francisco: San Francisco, CA 94158−2517, U.S.A., 2004.

(46) *OpenBabel, version 2.1.0*; The Open Source Chemistry Toolbox; SourceForge, 2007.

(47) *SMARTS*; A Language for Describing Molecular Patterns; Daylight, Inc.: Aliso Viejo, CA 92656, U.S.A., 1987.

(48) Bush, B. L.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.

(49) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(50) 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. D. S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT 06492, U.S.A., 2004.

(51) Sarkhel, S.; Desiraju, G. R. *Proteins* **2004**, *54*, 247–259.

(52) Kroon, J.; Kanters, J. A. *Nature* **1974**, *248*, 667–669.

(53) Scheiner, S. *Reviews in Computational Chemistry, Calculating the Properties of Hydrogen Bonds by Ab Initio Methods*; Wiley: Hoboken, NJ 07030, U.S.A., 1991; Vol. 2.

(54) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, *102*, 3782–3797.

(55) Lukin, O.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 6775–6782.

(56) Ireta, J.; Neugebauer, J.; Scheffler, M. *J. Phys. Chem. A* **2004**, *108*, 5692–5698.

(57) Kone, M.; Illien, B.; Graton, J.; Laurence, C. *J. Phys. Chem. A* **2005**, *109*, 11907–11913.

(58) Hao, M. H. *J. Chem. Theory Comput.* **2006**, *2*, 863–872.

(59) Raub, S.; Marian, C. M. *J. Comput. Chem.* **2007**, *28*, 1503–1515.

(60) Desiraju, G. R. *Acc. Chem. Res.* **1991**, *24*, 290–296.

(61) Taylor, R.; Kennard, O. *J. Am. Chem. Soc.* **1982**, *104*, 5063–5070.

(62) Derewenda, Z. S.; Lee, L.; Derewenda, U. *J. Mol. Biol.* **1995**, *252*, 248–262.

(63) Klaholz, B. P.; Moras, D. *Structure* **2002**, *10*, 1197–1204.

(64) Chakrabarti, P.; Chakrabarti, S. *J. Mol. Biol.* **1998**, *284*, 867–873.

(65) Pierce, A. C.; Sandretto, K. L.; Bemis, G. W. *Proteins* **2002**, *49*, 567–576.

(66) Panigrahi, S. K.; Desiraju, G. R. *Proteins* **2007**, *67*, 128–141.

(67) Jiang, L.; Lai, L. H. *J. Biol. Chem.* **2002**, *277*, 37732–37740.

(68) Madan Babu, M.; Kumar Singh, S.; Balaram, P. *J. Mol. Biol.* **2002**, *322*, 871–880.

(69) Baures, P. W.; Wiznycia, A.; Beatty, A. M. *Bioorg. Med. Chem.* **2000**, *8*, 1599–1605.

(70) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Am. Chem. Soc.* **2000**, *122*, 4750–4755.

(71) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Phys. Chem. A* **2000**, *104*, 5115–5121.

(72) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 4963–4968.

(73) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2005**, *109*, 6991–6992.

(74) Novoa, J. J.; Lafuente, P.; Mota, F. *Chem. Phys. Lett.* **1998**, *290*, 519–525.

(75) Yohannan, S.; Faham, S.; Yang, D.; Grosfeld, D.; Chamberlain, A. K.; Bowie, J. U. *J. Am. Chem. Soc.* **2004**, *126*, 2284–2285.

(76) Musah, R. A.; Jensen, G. M.; Rosenfeld, R. J.; McRee, D. E.; Goodin, D. B.; Bunte, S. W. *J. Am. Chem. Soc.* **1997**, *119*, 9083–9084.

(77) Toth, G.; Bowers, S. G.; Truong, A. P.; Probst, G. *Curr. Pharm. Des.* **2007**, *13*, 3476–3493.

(78) Sola, J.; Riera, A.; Verdaguer, X.; Maestro, M. A. *J. Am. Chem. Soc.* **2005**, *127*, 13629–13633.

(79) Scheiner, S.; Kar, T.; Gu, Y. *J. Biol. Chem.* **2001**, *276*, 9832–9837.

CT800267X

JCTC Journal of Chemical Theory and Computation

# Implementation of Umbrella Integration within the Framework of the Empirical Valence Bond Approach

Dhruva K. Chakravorty, Malika Kumarasiri, Alexander V. Soudackov, and Sharon Hammes-Schiffer*

*Department of Chemistry, 104 Chemistry Building, Pennsylvania State University, University Park, Pennsylvania 16802*

Received August 15, 2008

**Abstract:** The umbrella integration method for calculating the potential of mean force (PMF) for a chemical reaction is implemented within the empirical valence bond (EVB) framework. In this implementation, the PMF is generated along the energy gap reaction coordinate, and the biasing potential is the difference between the mapping potential, which is defined to be a linear combination of the valence bond state energies, and the EVB ground state energy. The umbrella integration method is based on the derivative of the PMF with respect to the reaction coordinate. An analytical expression for this derivative applicable to certain types of EVB potentials is presented. The advantages of the umbrella integration method are illustrated by the application of both umbrella integration and the weighted histogram analysis method to the hydride transfer reaction catalyzed by the enzyme dihydrofolate reductase. This application demonstrates that the umbrella integration method reduces the statistical errors, converges efficiently, and does not require significantly overlapping windows. A modified version of the weighted histogram analysis method that shares these advantages is also proposed and implemented.

## I. Introduction

The calculation of free energy barriers for chemical reactions is critical for predicting reaction rates. The free energy barrier is typically obtained by generating the potential of mean force (PMF) along a specified reaction coordinate. In umbrella sampling,[1] the PMF is generated by performing molecular dynamics or Monte Carlo simulations with a series of biasing potentials that enable sampling of the entire relevant range of the reaction coordinate. The probability distribution along the reaction coordinate for each biasing potential is obtained using standard binning techniques. Various methods have been developed for combining the probability distributions for the different biasing potentials to obtain the complete PMF for the unbiased system. The weighted histogram analysis method (WHAM) has been used extensively for this purpose.[2−7] Recently, Kästner and Thiel presented the alternative umbrella integration (UI) method.[8,9] The advantages of the UI method are that it avoids the iterative procedure inherent to WHAM, reduces the statistical errors, and converges more efficiently.[8,9] The previous implementation of UI considered only biasing potentials in the form of harmonic restraints along the reaction coordinate.[8,9]

In this paper, we implement UI within the framework of the empirical valence bond (EVB) approach, in conjunction with an energy gap reaction coordinate and nonharmonic biasing potentials defined in terms of mapping potentials. The EVB approach has been used successfully to describe a wide range of chemical reactions in solution and proteins.[10−14] In this approach, the chemical reaction is described in terms of a small number of valence bond states, and the EVB electronic ground state is obtained by diagonalizing the Hamiltonian matrix formed in the basis of these valence bond states. Single proton, hydride, and electron transfer reactions are often described in terms of two valence bond states, and the energy gap reaction coordinate is defined to be the difference between the energies of these two valence bond states. When umbrella sampling is used to generate the PMF along the energy gap reaction coordinate, the biasing potential may be chosen to be the energy difference between a mapping potential, which is a linear

* Corresponding author e-mail: shs@chem.psu.edu.

10.1021/ct8003386 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/08/2008

combination of the energies of the two valence bond states, and the EVB electronic ground state energy. Previously, we used thermodynamic integration and WHAM to generate the PMF within the framework of this EVB approach for charge transfer reactions in enzymes.[15−18] We also proposed and utilized an approach for calculating the rate constant from this PMF.[19] The implementation of UI within this framework provides an alternative method with the advantages enumerated above.

An outline of the paper is as follows. In Section II, we summarize the WHAM and UI approaches and present the equations required for the implementation of UI within the framework of the EVB approach. In Section III, we use both WHAM and UI to generate the PMF for the hydride transfer reaction catalyzed by the enzyme dihydrofolate reductase (DHFR). Our analysis of these calculations illustrates the advantages of UI over WHAM for this type of system. We also propose and implement a modification to WHAM that leads to similar advantages. The conclusions are presented in Section IV.

## II. Methods

In umbrella sampling,[1] simulations are performed with a series of biasing potentials $w_i(\xi)$, where $\xi$ is the reaction coordinate. The distribution $P_i^{\mathrm{b}}(\xi)$ of the biased system along the reaction coordinate is typically obtained by standard binning procedures to generate a histogram. Specifically, the relevant range of the reaction coordinate is divided into bins, and $P_i^{\mathrm{b}}(\xi_{\mathrm{bin}})$ is the fraction of sampled configurations in the bin centered at the reaction coordinate $\xi_{\mathrm{bin}}$ for the window corresponding to the biasing potential $w_i(\xi)$. The PMF for the biased system along the reaction coordinate is given by

$$A_i^{\mathrm{b}}(\xi) = -\frac{1}{\beta}\ln P_i^{\mathrm{b}}(\xi) \tag{1}$$

where $\beta = 1/k_BT$. The PMF for the unbiased system in each window is

$$A_i^{\mathrm{u}}(\xi) = -\frac{1}{\beta}\ln P_i^{\mathrm{b}}(\xi) - w_i(\xi) + F_i \tag{2}$$

where $F_i$ are constants that differ for each biasing potential or window.

In WHAM,[2−6] the constants $F_i$ are calculated iteratively to combine the unbiased potentials of mean force for different windows. The following two equations are solved iteratively:

$$P(\xi) = \sum_i^{\text{windows}} N_i P_i^{\mathrm{b}}(\xi) \Big/ \sum_j^{\text{windows}} N_j e^{[F_j - w_j(\xi)]\beta} \tag{3}$$

$$e^{-F_i\beta} = \int d\xi e^{-w_i(\xi)\beta} P(\xi) \tag{4}$$

where $N_i$ is the total number of configurations sampled for window $i$ used to construct $P_i^{\mathrm{b}}(\xi)$. After these equations are solved to self-consistency, the PMF $A(\xi)$ is obtained directly from $P(\xi)$ using the relation $A(\xi) = -\ln P(\xi)/\beta$.

In UI,[8,9] the derivative of the unbiased PMF with respect to the reaction coordinate is calculated for each window:

$$\frac{\partial A_i^{\mathrm{u}}(\xi)}{\partial \xi} = -\frac{1}{\beta}\frac{\partial \ln P_i^{\mathrm{b}}(\xi)}{\partial \xi} - \frac{dw_i(\xi)}{d\xi} \tag{5}$$

The data from different windows are combined according to a weighted average:

$$\frac{\partial A(\xi)}{\partial \xi} = \sum_i^{\text{windows}} p_i(\xi)\left(\frac{\partial A_i^{\mathrm{u}}(\xi)}{\partial \xi}\right) \tag{6}$$

where

$$p_i(\xi) = N_i P_i^{\mathrm{b}}(\xi) \Big/ \sum_i^{\text{windows}} N_i P_i^{\mathrm{b}}(\xi) \tag{7}$$

Subsequently, $A(\xi)$ is obtained by numerical integration over $\xi$. In previous applications of UI, the biasing potential is assumed to be of the form $w_i(\xi) = K(\xi - \xi_i)^2/2$. Moreover, the biased PMF is expanded in a power series and truncated after the quadratic term, which is equivalent to assuming a normal distribution for $P_i^{\mathrm{b}}(\xi)$:

$$P_i^{\mathrm{b}}(\xi) = \frac{1}{\sigma_i^{\mathrm{b}}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right)^2\right] \tag{8}$$

where the mean $\overline{\xi_i^{\mathrm{b}}}$ and the variance $\sigma_i^{\mathrm{b}}$ for each window are determined from the simulation data. These approximations lead to an analytical expression for the derivative of the unbiased PMF given in eq 5.

The UI method differs from WHAM in two important aspects. First, the UI method is based on the derivative of the PMF, rather than the PMF itself, so it does not involve offsets and therefore avoids the iterative procedure inherent to WHAM. Second, UI does not require a binning procedure because the mean and variance of the normal distribution for each window are determined directly from the raw simulation data, so a binning procedure is not required to obtain the derivative of the PMF given in eq 6. Specifically, the values of the reaction coordinate for all configurations sampled are collected during the simulation, and the mean and variance of the reaction coordinates collected for each window are determined directly from these data without generating a histogram. Moreover, in our implementation, the numerical integration of this derivative to generate the PMF is performed using an adaptive integration method that is converged to a specified precision without requiring the specification of a bin width. These numerical integrals are evaluated using the global adaptive strategy[20] in conjunction with the Gauss-Kronrod quadrature rule[21] as implemented in the Mathematica software package.[22]

To facilitate a meaningful comparison of the WHAM and UI methods, we propose a modified version of the WHAM method, denoted WHAM(n), that also avoids the binning procedure. In WHAM(n), the biased distribution $P_i^{\mathrm{b}}(\xi)$ for each window is represented by the normal distribution given in eq 8, where the mean and variance of $\xi$ for each window are determined directly from the simulation data. The WHAM equations given in eqs 3 and 4 are still solved iteratively, but $P_i^{\mathrm{b}}(\xi)$ in eq 3 is represented by the analytical normal distribution rather than the histogram obtained from a binning procedure. The integration in eq 4 is performed numerically using the adaptive integration method discussed above, thereby eliminating the necessity of specifying a bin width. Statistical methods[23] may be used to determine the

error bars for the mean and variance of $\xi$ used in eq 8 for both UI and WHAM(n). In addition to these statistical errors, a truncation error is introduced for both of these methods due to the approximation of the biased distribution by a normal distribution. A detailed analysis of the different sampling errors associated with UI has been performed for an analytical model potential.[9]

The main objective of this paper is to implement the UI method within the framework of a two-state EVB potential using an energy gap reaction coordinate and a mapping potential. For a two-state EVB model, the ground state EVB energy is

$$V_{\mathrm{EVB}} = \frac{1}{2}(V_{11} + V_{22}) - \frac{1}{2}\sqrt{(V_{11} - V_{22})^2 + 4V_{12}^2} \qquad (9)$$

where $V_{11}$ and $V_{22}$ are the energies of VB states 1 and 2, respectively, and $V_{12}$ is the coupling between these two states. In general, all of these quantities depend on the nuclear coordinates of the system. The energy gap reaction coordinate is defined as $\xi = V_{11} - V_{22}$. The simulations are performed with mapping potentials

$$V^i_{\mathrm{map}} = (1 - \lambda_i)V_{11} + \lambda_i V_{22} \qquad (10)$$

where the mapping parameter $\lambda_i$ is varied from zero to unity. The biasing potential is then of the form

$$w_i(\xi) = V^i_{\mathrm{map}} - V_{\mathrm{EVB}} = \left(\frac{1}{2} - \lambda_i\right)\xi + \frac{1}{2}\sqrt{\xi^2 + 4V_{12}^2} \qquad (11)$$

Note that this biasing potential is a function of only $\xi$ if $V_{12}$ is a function of only $\xi$. In this paper, we assume that $V_{12}$ is a constant, although the extension to the case in which $V_{12}$ is a function of $\xi$ is straightforward. Using this form for the biasing potential, the derivative of the unbiased PMF given in eq 5 is expressed as

$$\frac{\partial A_i^{\mathrm{u}}}{\partial \xi} = -\frac{1}{\beta}\frac{\partial \ln P_i^{\mathrm{b}}(\xi)}{\partial \xi} - \left(\frac{1}{2} - \lambda_i + \frac{\xi}{2\sqrt{\xi^2 + 4V_{12}^2}}\right) \qquad (12)$$

Approximating $P_i^{\mathrm{b}}(\xi)$ by a normal distribution, we have obtained an analytical form for the derivative of the unbiased PMF for each window. The data for the different windows can be combined using eq 6, followed by numerical integration of the derivative of the PMF over $\xi$ to obtain the PMF $A(\xi)$.

We also explore the use of different forms for the biased distribution $P_i^{\mathrm{b}}(\xi)$ because the mapping potential could lead to deviations from a normal distribution. We present results for the Gram-Charlier and the asymptotic Edgeworth expansions, which are expansions in terms of Chebyshev-Hermite polynomials. The Gram-Charlier expansion is of the form[24]

$$P_i^{\mathrm{b}}(\xi) = \frac{1}{\sigma_i^{\mathrm{b}}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right)^2\right] \times \left[1 + \frac{\kappa_3}{3!\,(\sigma_i^{\mathrm{b}})^5}He_3\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right) + \frac{\kappa_4}{4!\,(\sigma_i^{\mathrm{b}})^6}He_4\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right) + \frac{10}{6!}\left(\frac{\kappa_3}{(\sigma_i^{\mathrm{b}})^5}\right)^2 He_6\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right) + \ldots\right] \qquad (13)$$

where

$$He_n(x) = (-1)^n e^{x^2/2}\frac{d^n}{dx^n}e^{-x^2/2}$$

are Chebyshev-Hermite polynomials, $\sigma_i^{\mathrm{b}}$ is the variance, and $\kappa_n$ are the cumulants of the distribution $P_i^{\mathrm{b}}(\xi)$. The asymptotic Edgeworth expansion can be presented in the following compact form[25]

$$P_i^{\mathrm{b}}(\xi) = \frac{1}{\sigma_i^{\mathrm{b}}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\xi - \overline{\xi_i^{\mathrm{b}}}}{\sigma_i^{\mathrm{b}}}\right)^2\right]\left\{1 + \sum_{s=1}^{\infty}(\sigma_i^{\mathrm{b}})^s \times \sum_{\{k_m\}} He_{s+2r}(x)\prod_{m=1}^{s}\frac{1}{k_m!}\left(\frac{S_{m+2}}{(m+2)!}\right)^{k_m}\right\} \qquad (14)$$

where $S_n \equiv \kappa_n/(\sigma_i^{\mathrm{b}})^{2n-2}$, $\{k_m\}$ are the solutions of the Diophantine equation $k_1 + 2k_2 + \ldots + sk_s = s$, and $r = k_1 + k_2 + \ldots + k_s$. These asymptotic expansions are useful when the biased distributions for some windows differ from the normal distributions. The derivatives of the asymptotic expansions can still be evaluated analytically, and the moments and cumulants of the biased distributions can be calculated directly from the raw sampling data.

## III. Application

We use both WHAM and UI to calculate the PMF for hydride transfer in the enzyme DHFR. In this reaction, the hydride is transferred from the NC4 position of the NADPH (nicotinamide adenine dinucleotide phosphate) cofactor to the C6 position of the protonated dihydrofolate substrate. This reaction is depicted in Figure 1. We studied this reaction previously with a hybrid quantum-classical molecular dynamics approach, which includes the nuclear quantum effects of the transferring hydrogen with grid-based or path integral methods.[16−18] Here we use the same simulation system and EVB potential but do not include the nuclear quantum effects for simplicity. Since the simulation details are given elsewhere,[16,17] we provide only a brief summary in the present paper.

The simulation system includes the entire protein, the substrate, and the cofactor solvated by 4122 explicit water molecules in a truncated octahedral periodic box. The initial coordinates were obtained from a crystal structure of *Escherichia coli* DHFR complexed with $NADP^+$ and folate (PDB code 1rx2).[26] The potential energy surface is represented by a two-state EVB potential,[10] where state 1 corresponds to the transferring hydrogen atom bonded to the donor, and state 2 corresponds to the transferring hydrogen atom bonded to the acceptor. The diagonal elements of the EVB Hamiltonian are based on the GROMOS force field[27] with the EVB parameters given in ref 17. The two EVB parameters corresponding to the relative energy of the two valence bond states and the coupling between these states were fit to the experimental free energies of reaction and activation.[28]

In previous simulations, we used a set of 20 mapping parameters and performed 4.5 ns of molecular dynamics for each window with an additional 2 ns for the four windows near the transition state.[17] For the analysis in the present paper, we generated new data, starting with a snapshot from

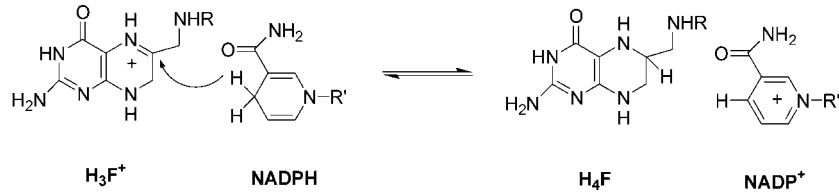



***Figure 1.*** Hydride transfer reaction from the NADPH cofactor to the protonated dihydrofolate substrate $H_3F^+$ to form the products tetrahydrofolate $H_4F$ and $NADP^+$.

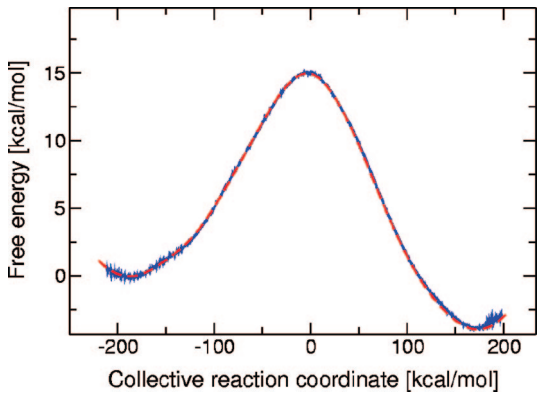



***Figure 2.*** PMF for the hydride transfer reaction generated with UI (red dashed) and WHAM (blue solid) with a bin size of 1.0 kcal/mol.

a reactant window in the previous simulation. We used a set of 19 mapping parameters from $\lambda_i = 0.05$ to 0.95 with a spacing of 0.05. The starting configuration for each window was obtained from the previous window after 20 ps of equilibration. Each window was equilibrated for a total of 350 ps, followed by 300 ps of data collection. We also generated two other independent sets of data with 50 ps of equilibration followed by 300 ps of data collection. The free energy barriers determined from these three data sets, as well as the previous simulations,[17,18] differ by less than 0.5 kcal/mol.

Figure 2 illustrates that the PMF curves generated with UI and WHAM are very similar. The free energy barriers of 15.0 and 15.3 kcal/mol determined with UI and WHAM, respectively, are consistent with the classical barriers determined from previous simulations using both thermodynamic integration and WHAM. However, the WHAM curve exhibits more numerical noise, particularly in the reactant and product wells. The WHAM curve in Figure 2 was generated with a bin size of 1 kcal/mol. The impact of bin size on the systematic and statistical errors in WHAM has been discussed in the literature.[29]

As discussed above, an advantage of UI is that it does not require a binning procedure for the simulation data, although it does require numerical integration to generate the PMF from its derivative. In contrast, WHAM relies on a binning procedure to generate the biased distributions used in the iterative procedure to determine the overall unbiased distribution. Moreover, WHAM does not converge as the number of bins increases (i.e., as the bin width decreases) because the statistical error increases as the number of bins increases.[29] In particular, the bin width must be sufficiently large to ensure that a sufficient number of configurations are sampled for each bin. Insufficient sampling per bin leads to

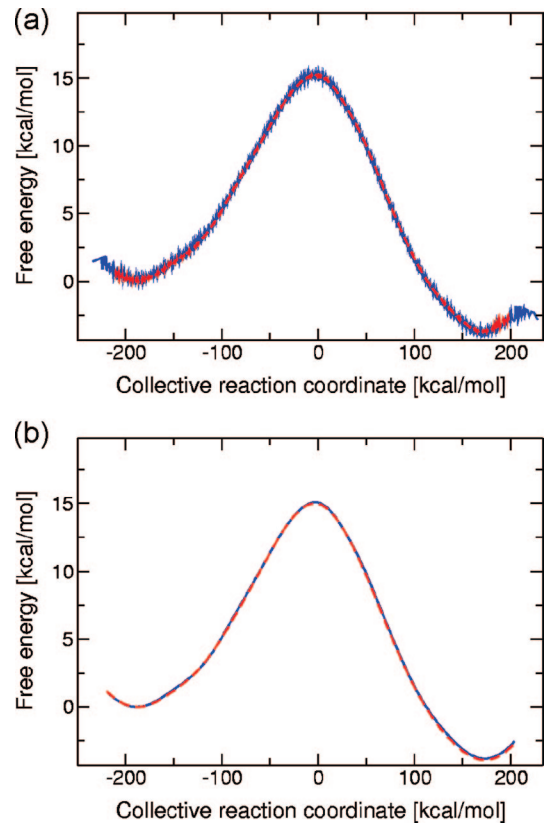



***Figure 3.*** PMF for the hydride transfer reaction generated with (a) WHAM using a bin size of 1.0 kcal/mol (red dashed) and 0.2 kcal/mol (blue solid) and (b) UI (red dashed) and WHAM(n) (blue solid). The UI and WHAM(n) PMF curves are virtually indistinguishable.

large statistical fluctuations that can result in substantial inaccuracies in the probability densities generated with WHAM. These difficulties with statistical error are avoided in UI because the biased distribution is represented by the analytical normal distribution function given in eq 8, where the mean and variance of the reaction coordinate for each window are obtained directly from the simulation data. In addition, a low weight is assigned to the tails of the distribution from each window in UI, as indicated by eq 7. As mentioned above, statistical methods[23] may be used to provide well-defined error bars for the mean and variance of the reaction coordinate, which can be propagated to estimate the sampling error for the resulting PMF.

Figure 3a illustrates the impact of bin size on the PMF curve generated with WHAM. Decreasing the bin size from 1.0 to 0.2 kcal/mol significantly increases the statistical noise of the PMF generated with WHAM. For comparison, Figure 3b depicts the PMF curve generated with the WHAM(n) method. This figure indicates that the statistical errors in

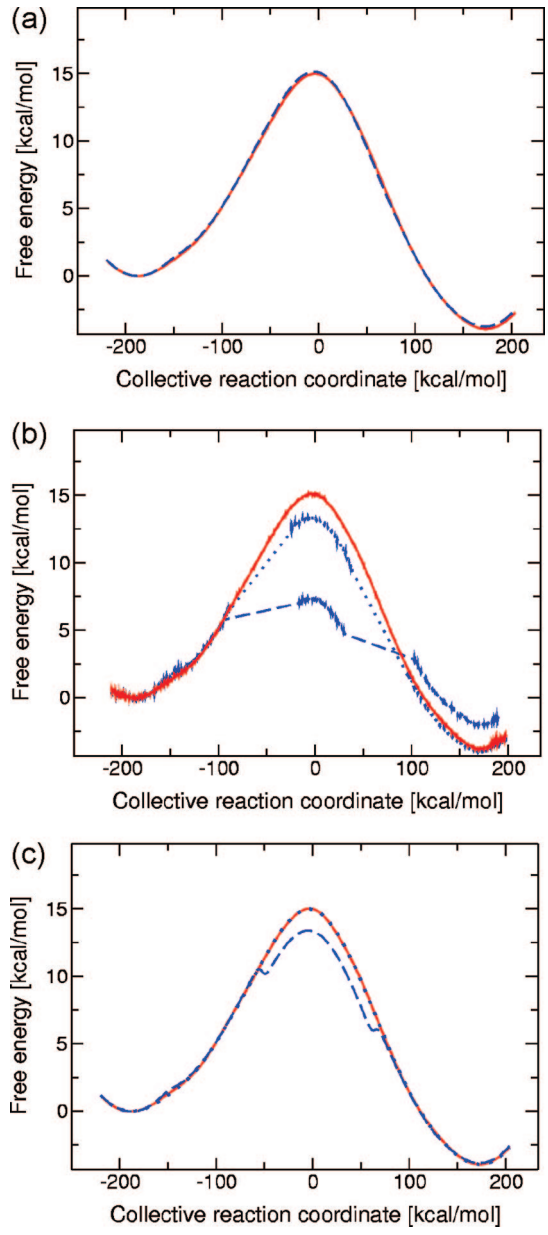



***Figure 4.*** PMF for the hydride transfer reaction generated using 19 windows (red solid) and five windows (blue dashed or dotted) with (a) UI, (b) WHAM, and (c) WHAM(n). For UI, the two PMF curves are virtually indistinguishable. The PMF curves generated with WHAM are shown for a convergence criterion of $10^{-4}$ (dashed) and $10^{-8}$ (dotted). The PMF curves generated with WHAM(n) are shown for a convergence criterion of $10^{-4}$ (dashed) and $10^{-6}$ (dotted). The 19 windows correspond to equally spaced values of $\lambda_i$ in the range $\lambda_i =$ 0.05 to 0.95, and the five windows correspond to $\lambda_i = 0.05$, 0.15, 0.50, 0.85, 0.95.

WHAM are significantly reduced when the biased distribution for each window is represented by the analytical normal distribution function rather than the histogram obtained from the binning procedure. This figure also illustrates that the PMF curve generated with WHAM(n) is virtually indistinguishable from the PMF curve generated with UI.

Another advantage of UI is that it does not require overlap between the distributions of the windows, although such overlap is desirable to enhance the accuracy. In contrast, WHAM requires sufficient overlap between the distributions of the windows. Figure 4 depicts the PMF generated with UI, WHAM, and WHAM(n) using only five windows corresponding to $\lambda_i = 0.05$, 0.15, 0.50, 0.85, and 0.95 (i.e., two windows in the reactant and product regions and one window in the barrier region). The PMF curve generated with UI using only five windows is virtually identical to the curve generated with all 19 windows. In contrast, the PMF curve generated with WHAM using only five windows is clearly problematic, as illustrated by Figure 4b. The barrier improves as the convergence criterion for the constants $F_i$ determined during the iterative procedure is tightened from a maximum change of $10^{-4}$ to $10^{-8}$, but the number of iterations required for convergence increases to more than $7.6 \times 10^7$ for a convergence criterion of $10^{-8}$, which still does not generate a smooth PMF. As shown in Figure 4c, the PMF curve generated with WHAM(n) using only five windows is better than that generated with WHAM for the same convergence criterion, but WHAM(n) still requires more than $8.6 \times 10^4$ iterations for a convergence criterion of $10^{-6}$, which generates a PMF that is indistinguishable from the PMF generated with WHAM(n) using all 19 windows.

In principle, given sufficient sampling within each window, WHAM and UI should converge to the same results if the distributions are Gaussian. However, the convergence of the iterative procedure in WHAM becomes slow for small overlap between the distributions of the windows, and insufficient sampling of the tail regions of the distributions combined with very small overlap could preclude convergence. An advantage of UI is that it utilizes an analytical expression for the distributions, thereby decreasing the statistical noise. Moreover, UI does not require an iterative procedure, so convergence is not an issue. These advantages become particularly pronounced for small overlaps between the distributions of the windows, although additional windows will enhance the accuracy of both methods.

Lastly, we test the approximation of the biased distribution function $P_i^{\mathrm{b}}(\xi)$ by a normal distribution, as given in eq 8. For this purpose, we explore the use of the Gram-Charlier and Edgeworth expansions.[25] The data and biased distribution functions for a representative window in the reactant region are shown in Figure 5a. The Gram-Charlier expansion is virtually indistinguishable from the normal distribution, whereas the Edgeworth expansion slightly improves the fit of the distribution obtained from the simulation data. As shown in Figure 5b, however, all three distribution functions lead to indistinguishable PMF curves. These data indicate that the approximation of the biased distribution by a normal distribution function is sufficient for generating quantitatively accurate PMF curves for this system. Note that this approximation may not be valid for certain systems, particularly when weak biasing potentials are used for free energy surfaces with high barriers or extended flat regions. In these cases, the WHAM method based on histograms obtained from a binning procedure could be more effective than the UI method.

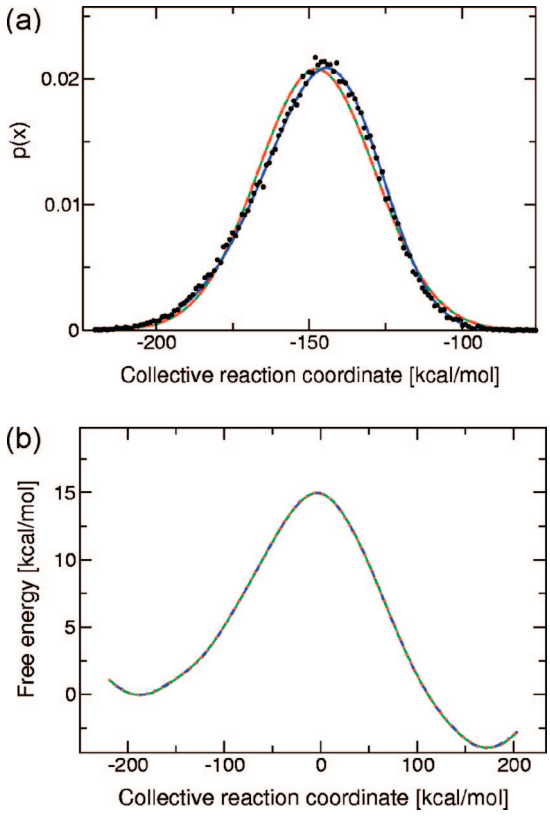



***Figure 5.*** (a) Biased probability distribution function for the window with $\lambda_i = 0.10$, where the filled circles represent the normalized histogram constructed from the simulation data and the solid and dashed lines represent fits to a normal distribution (red), a Gram-Charlier expansion to third order (green), and an Edgeworth expansion with three terms (blue). (b) PMF curves generated with the three fits in (a). The three PMF curves are virtually indistinguishable.

## IV. Conclusions

In this paper, we implemented the UI method for calculating the PMF along an energy gap reaction coordinate within the EVB framework. The UI method is based on the derivative of the PMF with respect to the reaction coordinate rather than the PMF itself. In this implementation, the biasing potential is the difference between the mapping potential, which is defined to be a linear combination of the valence bond state energies, and the EVB ground state energy. This biasing potential can be expressed as an analytical function of the energy gap reaction coordinate for a two-state EVB model in which the coupling between the two states is constant or is a function of the reaction coordinate. In this case, the derivative of the biasing potential with respect to the reaction coordinate can be expressed analytically, and the implementation of the UI method is straightforward.

We applied the UI and WHAM methods to the hydride transfer reaction catalyzed by DHFR. We showed that the UI and WHAM methods generate very similar PMF curves, although the PMF curve generated with UI exhibited less statistical noise. We also showed that the representation of the biased probability distributions as normal distributions is reasonable by comparison to expansions including non-Gaussian effects. Furthermore, our analysis illustrated two significant advantages of UI over WHAM. The first advantage is that UI does not rely on a binning procedure to generate histograms and therefore reduces the statistical error and converges efficiently. We proposed a modified version of WHAM that shares this advantage by representing the biased probability distribution for each window as an analytical normal distribution function rather than the histogram obtained from a binning procedure. The second advantage is that UI can provide accurate PMF curves efficiently even with a small number of windows that do not overlap significantly. In this case, the modified version of WHAM can also provide accurate PMF curves but is more computationally expensive because it requires a large number of iterations for convergence. Thus, UI is a promising method for generating accurate PMF curves for large systems for which sampling may be limited.

**Acknowledgment.** We are grateful for support from NIH grant GM56207.

### References

(1) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(2) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635–2638.

(3) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(4) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(5) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(6) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.

(7) Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, *18*, 1450–1462.

(8) Kastner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.

(9) Kastner, J.; Thiel, W. *J. Chem. Phys.* **2006**, *124*, 234106.

(10) Warshel, A., *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons, Inc.: New York, 1991.

(11) Schmitt, U. W.; Voth, G. A. *J. Phys. Chem. B* **1998**, *102*, 5547–5551.

(12) Vuilleumier, R.; Borgis, D. *Chem. Phys. Lett.* **1998**, *284*, 71–77.

(13) Cembran, A.; Gao, J. *Theor. Chem. Acc.* **2007**, *118*, 211–218.

(14) Truhlar, D. G. *J. Comput. Chem.* **2007**, *28*, 73–86.

(15) Billeter, S. R.; Webb, S. P.; Iordanov, T.; Agarwal, P. K.; Hammes-Schiffer, S. *J. Chem. Phys.* **2001**, *114*, 6925–6936.

(16) Agarwal, P. K.; Billeter, S. R.; Hammes-Schiffer, S. *J. Phys. Chem. B* **2002**, *106*, 3283–3293.

(17) Wong, K. F.; Watney, J. B.; Hammes-Schiffer, S. *J. Phys. Chem. B* **2004**, *108*, 12231–12241.

(18) Wang, Q.; Hammes-Schiffer, S. *J. Chem. Phys.* **2006**, *125*, 184102.

(19) Watney, J. B.; Soudackov, A. V.; Wong, K. F.; Hammes-Schiffer, S. *Chem. Phys. Lett.* **2006**, *418*, 268–271.

(20) Malcolm, M. A.; Simpson, R. B. *ACM Trans. Math. Software* **1975**, *1*, 129–146.

(21) Piessens, R.; Branders, M. *Math. Comput.* **1974**, *28*, 135–139.

(22) *Mathematica, Version 6.0*; Wolfram Research, Inc.: Champaign, IL, 2007.

(23) Schiferl, S. K.; Wallace, D. C. *J. Chem. Phys.* **1985**, *83*, 5203–5209.

(24) Stuart, A.; Ord, J. K. *Kendall's Advanced Theory of Statistics*, 6th ed.; John Wiley & Sons, Inc.: New York, 1994; Vol. 1. Distribution Theory, pp 228−229.

(25) Blinnikov, S.; Moessner, R. *Astron. Astrophys. Suppl.Ser.* **1998**, *130*, 193–205.

(26) Sawaya, M. R.; Kraut, J. *Biochemistry* **1997**, *36*, 586–603.

(27) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular simulation: The GROMOS96 manual and user guide*; VdF Hochschulverlag ETH Zurich: Zurich, 1996.

(28) Fierke, C. A.; Johnson, K. A.; Benkovic, S. J. *Biochemistry* **1987**, *26*, 4085–92.

(29) Kobrak, M. N. *J. Comput. Chem.* **2003**, *24*, 1437–1446.

CT8003386

JCTC Journal of Chemical Theory and Computation

# Theoretical and Crystallographic Study of the Dual $\sigma/\pi$ Anion Binding Affinity of Quinolizinylium Cation

Carolina Estarellas, Antonio Frontera,* David Quiñonero, and Pere M. Deyà*

*Departament de Química, Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain*

Received August 12, 2008

**Abstract:** Benzoquinolizinylium salts are important compounds in the regulation of transmembrane conductance regulator channels. In this manuscript, the geometrical position ($\sigma/\pi$) of the counteranion in quinolizinylium salts has been studied by means of ab initio calculations at the RI-MP2(full)/6−31++G** level of theory. A search in the Cambridge Structural Database determines that the position of the anion depends upon its nature. Halogen anions prefer anion−$\sigma$ interactions, and $BF_4^-$ and $PF_6^-$ anions prefer anion−$\pi$ interactions. The dual $\sigma/\pi$ binding affinity of title compound has been studied by means of ab initio and molecular interaction potential with polarization (MIPp) calculations and the Bader's theory of "atoms-in-molecules".

## I. Introduction

Noncovalent interactions play a decisive role in many areas of modern chemistry. This is especially true in the field of supramolecular chemistry and molecular recognition.[1] Interactions involving aromatic rings are important binding forces in both chemical and biological systems, and they have been reviewed by Meyer et al.[2]

At the same time, Mascal et al.,[3] Alkorta et al.,[4] and our group[5] have demonstrated, theoretically, that the $\pi$-interaction of anions with electron deficient aromatic rings is favorable. A pioneering manuscript describing gas-phase clustering reactions between anions and hexafluorobenzene using both theoretical and experimental results was previously reported.[6] Our group has used the term "anion−$\pi$ interaction"[7] to describe the interaction between anions and hexafluorobenzene, where the anion is positioned over the ring along the $C_6$ axis.[5] The anion−$\pi$ interaction is dominated by electrostatic and anion-induced polarization terms.[4,5] The strength of the electrostatic component depends upon the value of $Q_{zz}$, and the anion induced polarization term correlates with the molecular polarizability ($\alpha_{||}$) of the aromatic compound.[8] Anion−$\pi$ complexes have been observed experimentally, sustaining the theoretical predictions and the promising proposal for the use of anion receptors based on anion−$\pi$ interactions in molecular recognition.[9−13] These interactions

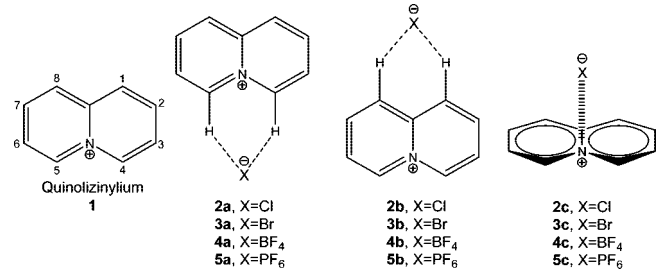



***Figure 1.*** Quinolizinylium cation **1** and its anion−$\pi/\sigma$ complexes **2**−**5**.

are also important in ADN bases, such as adenine.[14] Moreover, Berryman et al. have reported structural criteria for the design of anion receptors based on the interaction of halides with electron-deficient arenes.[15] Recent excellent reviews deal with anion-binding involving $\pi$-acidic heteroaromatic rings.[16]

Chloride channels play important roles in homeostasis and regulate cell volume, transepithelial transport, and electrical excitability.[17] The cystic fibrosis transmembrane conductance regulator (CFTR) is a cAMP-regulated epithelial chloride channel, mutations in which cause cystic fibrosis.[18] This syndrome is the most common lethal autosomal recessive genetic disease in caucasians. It has been demonstrated that benzoquinolizinylium chloride salts activate both wild-type and mutant cystic fibrosis transmembrane conductance regulator channels.[19] In this manuscript, we report a computational study, where we analyze the geometrical and

* To whom correspondence should be addressed. E-mail: toni.frontera@uib.es.

10.1021/ct800332y CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/10/2008

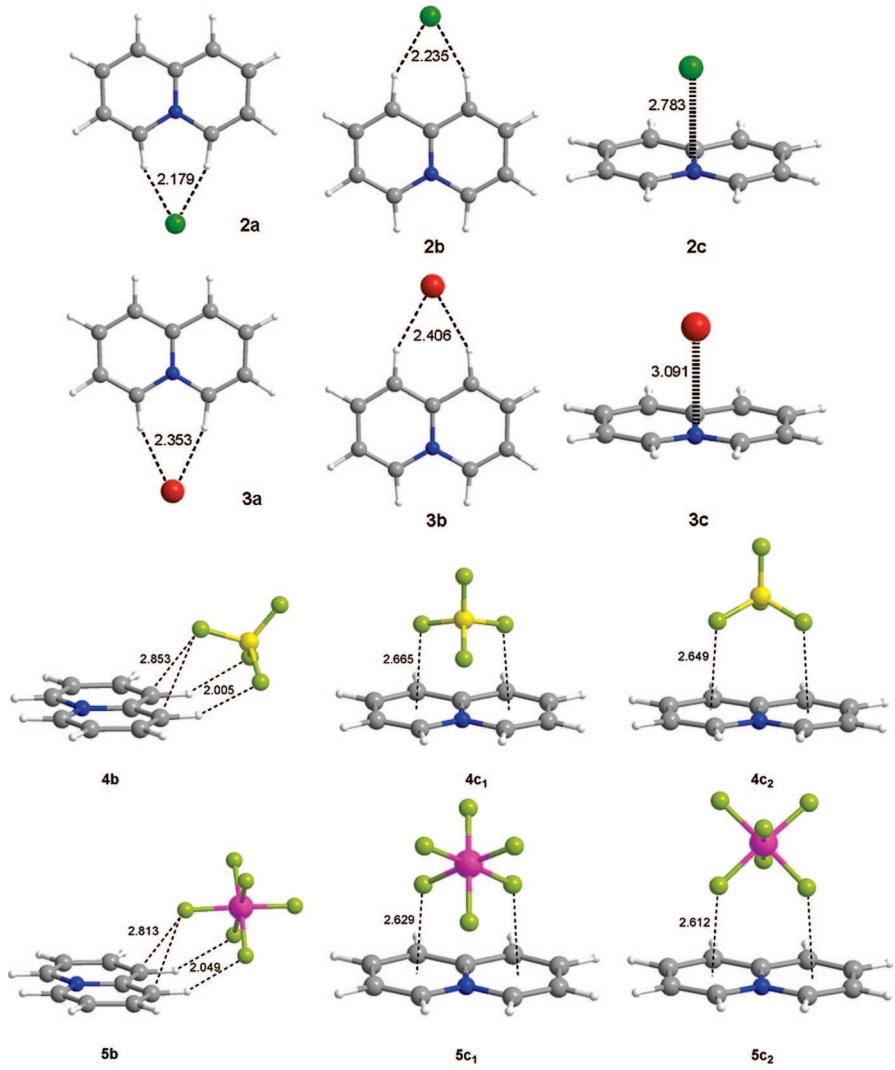



***Figure 2.*** RI-MP2/6−31++G** optimized complexes **2**−**5**. Distances in Å.

energetic features of anion−$\pi$ and anion−$\sigma$ complexes of several anions with quinolizinylium cation **1**, also known as quinolizinium. It has been recently reported a combined experimental and theoretical study that deals with the structural criteria for the design of anion receptors based on electron-deficient arenes.[15] The presence of electron-withdrawing groups (EWG) increases the acidity of the arene C−H donors, and consequently, it has a double effect. First, the anion can interact via hydrogen bonding with the arene because the presence of EWG strengthens the C−H···X$^-$ interaction. Second, the anion can also interact with the $\pi$-cloud of the arene because the presence of EWG increases the $\pi$-acidity of the ring. Cationic aromatic rings can be considered as an extreme of this situation. The aromatic ring is obviously electron-deficient because of its cationic character. In addition, all hydrogen atoms of the ring are available for hydrogen bonding because they have not been substituted by EWGs and their acidity is increased with respect to neutral aromatic rings. Recently, Alkorta el al. have studied a related system, the case of pentazolo[1,2-*a*]pentazole, ($N_8$) with neutral electron donors, hydrogen-bond donors, and anions.[20] The quinolizinylium cation can interact with the counterion either via the C−H aromatic groups ($\sigma$ interaction, i.e., hydrogen-bonded complex) or via an anion−$\pi$ interaction. This dual $\sigma/\pi$ binding affinity of **1** has been studied by means of ab initio and molecular interaction potential with polarization (MIPp) calculations, and the Bader's theory of "atoms-in-molecules".

Energetically, in relation to the cation−$\pi$ interaction, the anion−$\pi$ interaction is less favorable because the van der Waals radii of anions are bigger than cations, and consequently, the equilibrium distances are larger in anion−$\pi$ complexes than in cation−$\pi$ complexes.[21]The energetic terms that contribute to the stabilization of ion−$\pi$ interactions (electrostatic and polarization) are very dependent on the distance. Therefore, the potential use of electron-deficient aromatic rings as building blocks for the construction of anion receptors[22] is handicapped with respect to the widely use of cation receptors based on cation−$\pi$ interactions. Inspired by guanidinium salts that combine both hydrogen bonding and electrostatic forces to bind anions, we have recently proposed that a possible solution to solve this disadvantage is the use of charged aromatic compounds that increment the anion-binding ability of the ring.[23] The quinolizinylium cation is an example. Exploring the Cambridge Structural Database (CSD), we have found compounds containing quinolizinylium exhibiting anion−$\pi/\sigma$ interactions in the solid state. These interactions have an active influence

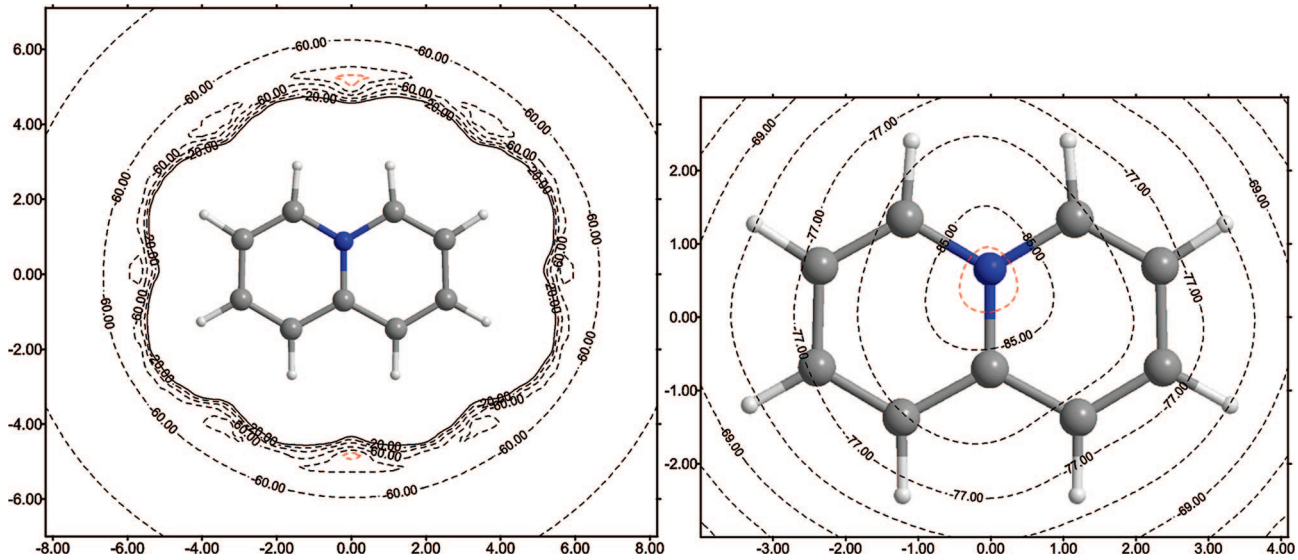


***Figure 3.*** Right: 2D-MIPp($Cl^-$) energy map computed for **1** at 2.8 Å above the molecular plane. Isocontour lines are plotted every 4 kcal/mol. The lowest energy isocontour is ploted in red. Axes units are Å, and energies are in kcal/mol. Left: 2D-MIPp($Cl^-$) energy maps computed for **1** at the molecular plane. Isocontour lines are plotted every 10 kcal/mol. The lowest energy isocontour is plotted in red.

***Table 1.*** Interaction Energies at the RI-MP2(full)/6-31++G** Level of Theory without and with the BSSE Corrections (E and $E_{BSSE}$, kcal/mol) and Equilibrium Distances ($R_e$, Å)[a]

| compound | NImag | $E$ | $E_{BSSE}$ | $R_e$ | $q$ (e) M-K | $q$ (e) Mull |
|---|---|---|---|---|---|---|
| **2a** | 0 | −92.02 | −86.23 | 2.179[b] | −0.82 | −0.72 |
| **2b** | 0 | −85.94 | −80.35 | 2.235[b] | −0.83 | −0.73 |
| **2c** | 0 | −89.42 | −81.56 | 2.783[c] | −0.77 | −0.70 |
| **3a** | 0 | −87.88 | −82.80 | 2.353[b] | −0.79 | −0.81 |
| **3b** | 0 | −82.43 | −77.38 | 2.406[b] | −0.82 | −0.78 |
| **3c** | 0 | −87.58 | −81.60 | 3.091[c] | −0.81 | −0.72 |
| **4a** | Stationary point not found, it converges to **4c** | | | | | |
| **4b** | 0 | −78.27 | −74.26 | 2.005[d] | −0.85 | −0.87 |
| **$4c_1$** | 0 | −86.38 | −79.30 | 3.220[e] | −0.78 | −0.89 |
| **$4c_2$** | 0 | −84.64 | −77.80 | 3.195[e] | −0.77 | −0.91 |
| **5a** | Stationary point not found, it converges to **5c** | | | | | |
| **5b** | 0 | −73.88 | −68.98 | 2.049[d] | −0.86 | −0.89 |
| **$5c_1$** | 0 | −82.57 | −74.31 | 3.662[e] | −0.80 | −0.92 |
| **$5c_2$** | 0 | −81.04 | −72.98 | 3.639[e] | −0.79 | −0.92 |

[a] The computed Merz−Kollman (M-K) and Mulliken (Mull) charges of the anion (q, e) are also included for both hydrogen-bonded and $\pi$-complexes. [b] $R_e$ is the mean distance of two C−H···$X^-$ distances. [c] $R_e$ is measured from the anion to the middle of the bridge C−N bond. [d] Measured from the F atom to the H atom of the ring as shown in Figure 2. [e] Measured from the B/P atom to the middle of the common C−N bond.

in the crystal packing of the quinolizinylium cation and the location of the anion agrees with the theoretical calculations.

## II. Theoretical Methods

The geometry of all the complexes included in this study was fully optimized at the RI-MP2/6−31++G** level of theory within the program TURBOMOLE, version 5.7.[24] The RI-MP2 method[25,26] applied to the study of cation−$\pi$ and anion−$\pi$ interactions is considerably faster than the MP2 and the interaction energies, and equilibrium distances are almost identical for both methods.[27,28] The binding energy was calculated at the same level with and without correction for the basis set superposition error (BSSE) using the Boys−Bernardi counterpoise technique.[29] No symmetry constraints have been imposed in the optimizations, and all compounds and complexes belong to the $C_1$ symmetry point group. Frequency calculations at the same level of theory have confirmed the minimum nature of all complexes.

The contributions to the total interaction energy have been computed using the molecular interaction potential with polarization (MIPp) methodology,[30] which is an improved generalization of the molecular electrostatic potential (MEP), where three terms contribute to the interaction energy: (i) an electrostatic term identical to the MEP,[31] (ii) a classical dispersion−repulsion term, and (iii) a polarization term derived from perturbation theory.[32] Calculation of the MIPp of **1** with $F^-$ and $Cl^-$ anions was performed using the HF/6−31++G**//RI-MP2(full)/6−31++G** wave function of the aromatic rings by means of the MOPETE-98 program.[33] Calculation of MIPp using the MP2 wave function are not available.[33] The ionic van der Waals parameters for $F^-$ and $Cl^-$ were taken from the literature.[34]

The topological analysis of the electron charge density performed for the complexes of **1** with anions was determined using Bader's theory of "atoms-in-molecules" (AIM).[35] The electronic density analysis was performed using the AIM2000 program[36] at the MP2//RI-MP2 level of theory. We have evaluated the charge transfer in the complexes by using the Merz−Kollman (M-K) scheme for deriving atomic charges at the MP2/6−31++G**//RI-MP2(full)/6−31++G** level of theory. It has been reported that this method provides high quality charges.[37] The Mulliken charges[38] are also included for comparison purposes.

## III. Results and Discussion

**A. Energetic and Geometrical Details.** Table 1 reports the energies and equilibrium distances corresponding to the interaction of **1** with several anions via either hydrogen bonding (**2a**−**5a** and **2b**−**5b**) or anion−$\pi$ (**2c**−**5c**) binding

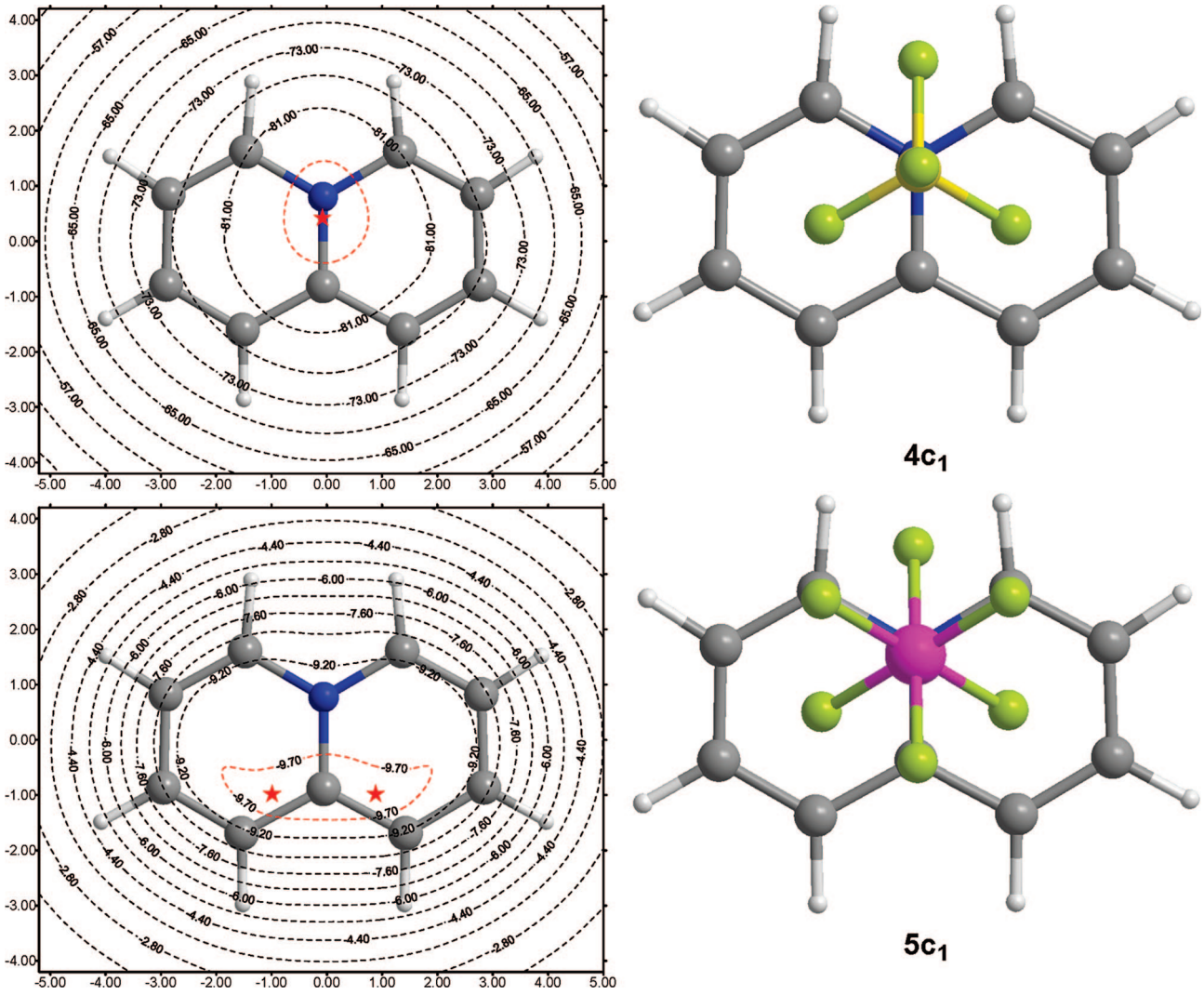



***Figure 4.*** Right: Zenithal views of the optimized anion−$\pi$ complexes **4c$_1$** and **5c$_2$**. Left, bottom: 2D-IPE($F^-$) energy map computed for **1** at 2.8 Å above the molecular plane. Isocontour lines are plotted every 0.8 kcal/mol. Left, top: 2D-MIPp($F^-$) energy map computed for **1** at 2.8 Å above the molecular plane. Isocontour lines are plotted every 4 kcal/mol. In both maps, the lowest-energy isocontour is plotted in red. Axes units are Å, and energies are in kcal/mol. The global minima are represented by red stars.

types. The geometry of the optimized complexes is depicted in Figure 2. For all complexes, the interaction energies are large and negative because of the ion−pair nature of the interaction. From the inspection of the results, several interesting points arise. First, the most favorable situation for the complexation of **1** with $Cl^-$ and $Br^-$ is the hydrogen bonding interaction via the hydrogen atoms that are in α with respect to the aromatic nitrogen atom (**2a** and **3a**). Moreover, the anion−$\pi$ complexes **2c** and **3c** are more favorable than the hydrogen bonding complexes **2b** and **3b**, indicating that the $\pi$-binding mode is more favorable than the $\sigma$-interaction (bifurcated hydrogen bond) with both $H^1$ and $H^8$ hydrogen atoms (see Figure 1 for the numbering of **1**). A differentiating feature between anion-$\pi$ complexes **2c** and **3c** is that the chloride anion lies approximately over the carbon atom of the C−N common bond whereas the bromide anion lies approximately over the nitrogen atom of the common bond. We have explored all possible binding modes between **1** and the $Cl^-$ anion through hydrogen bonding interactions, and in all cases, they converge to either **2a** or **2b**. The same result was obtained in the bromide complexes. Second, a different behavior is observed in the complexes of poliatomic anions $BF_4^-$ and $PF_6^-$. For both anions, the hydrogen bonding complexes **4a** and **5a** are not found; they converge to the anion−$\pi$ complexes **4c** and **5c**, respectively. This indicates that both anions prefer the $\pi$-binding mode with **1**. Moreover, for both anions, the $\pi$-complexes are considerably more favorable energetically than the hydrogen bonding complexes (**4b** and **5b**). For the anion−$\pi$ complexes **4c** and **5c**, two favorable orientations for the anion have been found, namely, **4c$_1$**, **4c$_2$**, **5c$_1$** and **5c$_2$**, see Figure 2 for details. Complexes **4c$_1$** and **5c$_1$** are slightly more favorable than **4c$_2$** and **5c$_2$**, respectively. Third, in Table 1, we also include the charge of the anion in the complexes to study charge transfer effects. We have used two methods for deriving the atomic charges, Mulliken and Merz−Kollman (M-K). It has been demonstrated that the latter method provides high-quality charges. It can be observed that in all complexes, the charge transfer (M-K) is important (about 0.2 e). Moreover, charge transfer is less important in hydrogen bonding complexes

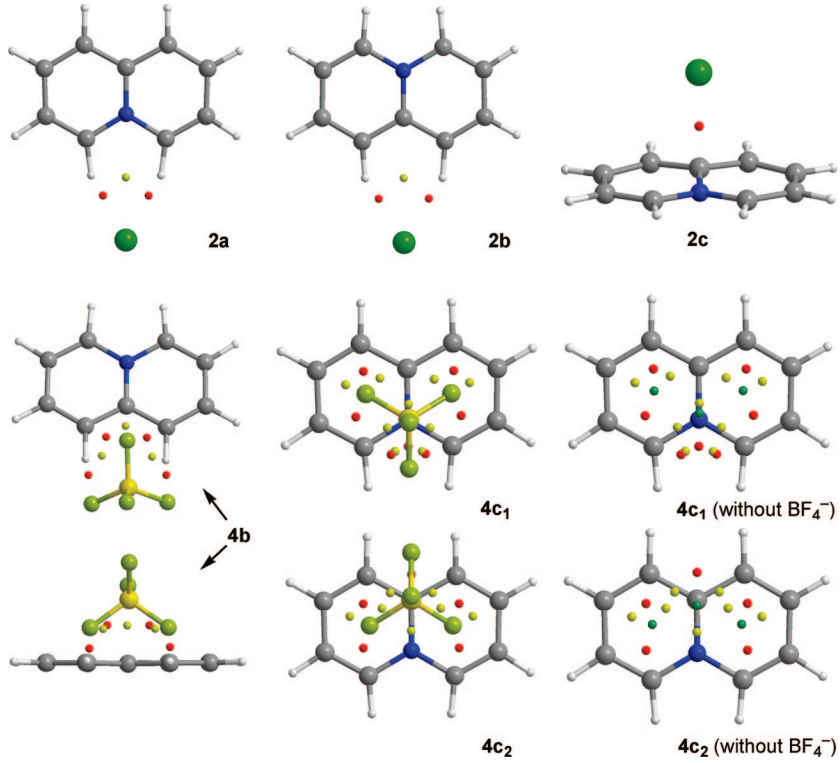



***Figure 5.*** Schematic representation of the critical points that emerge upon complexation of the anion for complexes **2a**−**2c** and **4b**−**4c**. Bond CPs are represented in red; ring CPs are in yellow, and cage CPs are in green.

***Table 2.*** Electron Density ($\rho$) and Its Laplacian ($\nabla^2\rho$) in Atomic Units at the Critical Points (CP) Originated upon Complexation and the Number ($n$) of each CP in the complex, Computed at the MP2/6-31++G**//RI-MP2/6-31++G** Level of Theory[a]

| compound | $n$ | CP | $10^2\rho$ | $10\nabla^2\rho$ |
|---|---|---|---|---|
| **2a** | 2 | (3, −1) | 2.780 | 0.706 |
| | 1 | (3, +1) | 1.207 | 0.581 |
| **2b** | 2 | (3, −1) | 2.473 | 0.662 |
| | 1 | (3, +1) | 0.906 | 0.459 |
| **2c** | 1 | (3, −1) | 2.594 | 0.704 |
| **3a** | 2 | (3, −1) | 2.369 | 0.564 |
| | 1 | (3, +1) | 1.107 | 0.494 |
| **3b** | 2 | (3, −1) | 2.212 | 0.535 |
| | 1 | (3, +1) | 0.833 | 0.392 |
| **3c** | 1 | (3, −1) | 1.602 | 0.522 |
| **4b** | 2 | (3, −1) | 2.057 | 0.733 |
| **4c$_1$** | 1 | (3, +3) | 0.622 | 0.379 |
| **4c$_2$** | 1 | (3, +3) | 0.568 | 0.362 |
| **5b** | 2 | (3, −1) | 1.835 | 0.677 |
| **5c$_1$** | 1 | (3, +3) | 0.561 | 0.397 |
| **5c$_2$** | 1 | (3, +3) | 0.515 | 0.342 |

[a] For complexes **4b** and **5b**, only the CPs that describe the $\sigma$-interaction are summarized. For complexes **4c** and **5c**, only the cage CPs that describe the interaction of the B/P with **1** are summarized.

than in $\pi$-complexes, apart from bromide complexes, probably because of the larger equilibrium distance of the anion−$\pi$ complex **3c** (3.091 Å). Mulliken charges do not follow the same trend. In some cases, they indicate a large charge transfer (0.3 e in **2c**), and in other cases, they indicate a very small charge transfer (0.08 e in **5c**). These inconsistencies confirm the low quality of the charges obtained from the Mulliken population analysis, in agreement with previous observations.[37]

The geometric features of complexes **2**−**5** are shown in Figure 2. The equilibrium distances of complexes **2a** and **3a** are shorter than the ones computed for **2b** and **3b**, respectively, in agreement with the energetic results. In the hydrogen-bonded complexes **2a**-**2b** and **3a**-**b**, the anion interacts simultaneously with two hydrogen atoms. Experimentally, it has been demonstrated that bifurcated hydrogen bonds to two neighboring CH groups are energetically favored over linear hydrogen bonds to a single CH groups.[39] The hydrogen bonding complexes **4b** and **5b** present a peculiar geometry, see Figure 2, where the anion is not coplanar with the quinolizinylium rings. Two fluorine atoms of the anion interact with two hydrogen atoms ($H^1$ and $H^8$), and another fluorine atom interacts with two carbon atoms of the ring ($C^1$ and $C^8$). Finally, in the anion−$\pi$ complexes, **4c$_{1-2}$** and **5c$_{1-2}$**, the anion is located approximately over the center of the bridge C−N bond. Two fluorine atoms point to the ring centroids and a third fluorine atom is located approximately over the nitrogen atom in **4c$_1$** and **5c$_1$** and over the carbon atom of the common bond in **4c$_2$** and **5c$_2$**.

**B. MIPp Analysis.** With the purpose of analyzing the nature of the anion−$\pi/\sigma$ interaction in the quinolizinylium cation and understanding the importance of electrostatic and polarization terms, we have performed the calculation of MIPp of **1** interacting with $F^-$ and $Cl^-$ using the HF/6−31++G**//RI-MP2(full)/6−31++G** wave function. In the calculations, the $F^-$ and $Cl^-$ ions were considered as a classical nonpolarizable particles. In Figure 3, we represent the bidimensional MIPp (2D-MIPp) maps obtained for **1** interacting with $Cl^-$. We have computed two 2D-MIPp maps, one at the molecular plane and the other at 3.0 Å over the molecular plane and parallel to it to study the anion-binding

***Table 3.*** Reference Codes of the X-ray Structures That Include the Quinolizinylium Moiety, the Counterion, the Type of Interaction and References

| CSD reference | entry | anion | interaction | ref |
|---|---|---|---|---|
| AZPHNE | 1 | $Cl^-$ | $\sigma$ | 47 |
| CIPQOH | 2 | $Br^-$ | $\sigma$ | 48 |
| CIPQUN | 3 | $Br^-$ | $\sigma$ | 48 |
| CIPROI | 4 | $BF_4^-$ | $\sigma$ and $\pi$ | 48 |
| DANJAD | 5 | $Br^-$ | $\sigma$ | 49 |
| DAVPOG | 6 | $BF_4^-$ | $\sigma$ and $\pi$ | 50 |
| DAVPUM | 7 | $BF_4^-$ | $\sigma$ and $\pi$ | 50 |
| GUKTEL | 8 | $PF_6^-$ | $\sigma$ | 51 |
| IDIFAD | 9 | $BF_4^-$ | $\sigma$ and $\pi$ | 52 |
| IXUJEQ | 10 | $Cl^-$ | $\sigma$ | 19 |
| IXUJIU | 11 | $Cl^-$ | $\sigma$ | 19 |
| IXUJOA | 12 | $Br^-$ | $\sigma$ | 19 |
| KASVOP10 | 13 | $PF_6^-$ | $\sigma$ and $\pi$ | 45 |
| KIHGOX | 14 | $BF_4^-$ | $\sigma$ and $\pi$ | 46 |
| LUCGEV | 15 | $Br^-$ | $\sigma$ | 53 |
| LUCGIZ | 16 | $Br^-$ | $\sigma$ | 53 |
| TATZOE | 17 | $CH_3SO_3^-$ | $\sigma$ and $\pi$ | 54 |
| TATZUK | 18 | $Cl^-$ | $\sigma$ | 54 |
| TAVBIC | 19 | $CH_3SO_3^-$ | $\sigma$ and $\pi$ | 54 |
| TAVBOI | 20 | $Cl^-$ | $\sigma$ | 54 |
| ULATIK | 21 | $Cl^-$ | $\sigma$ | 55 |
| WEPTEQ | 22 | $Br^-$ | $\sigma$ | 56 |
| WUXKAB | 23 | $PF_6^-$ | $\sigma$ and $\pi$ | 57 |

ability of **1**, first via hydrogen bonding using the hydrogen atoms of the ring and second via an anion-$\pi$ interaction, respectively. It can be observed a good agreement between the location of the minima in the 2D-MIPp energy maps and the geometry of the optimized complexes. For instance, it can be observed that in the 2D-MIPp($Cl^-$) energy map computed at the molecular plane several local minima are observed, each of them corresponds to the interaction of the anion with two C−H groups. The global MIPp minimum corresponds to the interaction of the anion with the two C−H groups α to the nitrogen atom with an interaction energy of −75.2 kcal/mol. An additional minimum, which is almost isoenergetic (−74.1 kcal/mol), is found at the opposite part of the map, as a result of the interaction of the anion with $H^1$ and $H^8$. The 2D-MIPp($Cl^-$) energy map computed at 3.0 Å above the molecular plane predicts the location of the anion over the common C−N bond, to some extent displaced toward the nitrogen atom. These results are in qualitative agreement with the ab initio calculations of complexes **2** and **3**. The main difference resides in the position of the chloride anion, which, in **2c**, is located over the C−N bond and somewhat displaced to the carbon atom and in the MIPp the minimum is displaced to the nitrogen atom.

We have recently demonstrated that the utility of a new tool to predict the geometries of anion--$\pi$ complexes, where the anion is polyatomic. This tool is entitled the induced-polarization energy map (IPE map). The novelty of this representation is that, in the map, only the contribution of the ion-induced polarization term to the total interaction energy is contoured in a 2D region. The IPE map has been found useful to predict and explain geometries of anion−$\pi$ complexes of tetrahedral $BF_4^-$ anion with several diazines, triazines, and tetrazines, and it nicely complements the MIPp map.[8b] We have computed the 2D-IPE($F^-$) of **1** at 2.8 Å above the molecular plane. The representation is shown in Figure 4 (left, bottom) together with the 2D-MIPp($F^-$) (left, top). The corresponding RI-MP2(full)/6−31++G** optimized complexes are also included in Figure 4 (right) to illustrate the agreement of the 2D-IPE/MIPp maps with the geometric features of the complexes. The IPE map indicates that there is a wide region (red contour) that includes both rings where the IPE energy is minimum. We have found two global IPE minima that are represented in the 2D map by red stars. The MIPp minimum is located approximately over the nitrogen atom (represented by a red star in the map). The MP2/6−31++G** optimized complexes $\mathbf{4c_1}$ and $\mathbf{5c_1}$ are in agreement with both maps. The global position of the anion is located where the MIPp map predicts with two fluorine atoms closely located at the lowest isocontour line of the IPE map. Therefore, the 2D-IPE and 2D-MIPp maps can be combined to predict and explain the observed geometric features of the optimized $\pi$-complexes. This agreement is most likely observed because the energetically important interactions described by the maps are those used to determine the geometry in the first place when electronic energies are minimized using electron-correlated ab initio methods. In the complexes $\mathbf{4c_1}$ and $\mathbf{5c_1}$, a direct comparison of the MIPp energy values and the MP2/6−31++G** interaction energies is not possible because the MIPp maps are computed using $F^-$ as the interacting particle instead of $BF_4^-$, because of limitations of the MOPETE-98 program.

**C. AIM Analysis.** Topological analysis of the charge density $\rho$(r) distribution and properties of critical points (CP) were determined for complexes **2**−**5** using the Bader's theory of "atoms-in-molecules", which provides an unambiguous definition of chemical bonding,[40] using the MP2(full)/6−31++G** wave function. The AIM theory has been successfully used to characterize anion−$\pi$ interactions.[5,7] For complexes **2a,b**−**3a,b**, the exploration of the CPs revealed the presence of two bond CPs that connect the anion with two hydrogen atoms. As a consequence of the geometry of the complexes, one ring CP is also generated. In Figure 5, we represent the distribution of CPs that are generated upon complexation of the anion in hydrogen bonding complexes **2a**−**b**, the distribution of CPs in complexes **3a**−**b** is identical, and they are not shown in the figure. For the anion−$\pi$ complexes **2c**−**3c**, the exploration of the CPs revealed the presence of only one bond CP that connects the anion with the carbon atom of the common C−N bond (see Figure 4, only **2c** is represented). The distribution of CPs in complexes **4**−**5** is more complicated. In hydrogen bonding complexes **4b**−**5b**, the exploration of CPs revealed the presence of four bond CPs and three ring CPs. Two bond CPs connect two fluorine atoms to two hydrogen atoms of the ring, and they properly describe the $\sigma$-interaction. The other CPs connect a third fluorine and the boron atoms with three carbon atoms of the ring. This can be considered as a pseudo-$\pi$-interaction. The distribution of CPs in the anion−$\pi$ complexes $\mathbf{4c_{1-2}}$ is very complicated. It can be viewed as a sum of three interactions: First is the interaction of two fluorine atoms of the $BF_4^-$ with both rings of **1**. Each of these two interactions is described by two bond, two ring and one cage CPs. The bond and ring CPs connect the fluorine atom with four carbon atoms of the ring. The cage

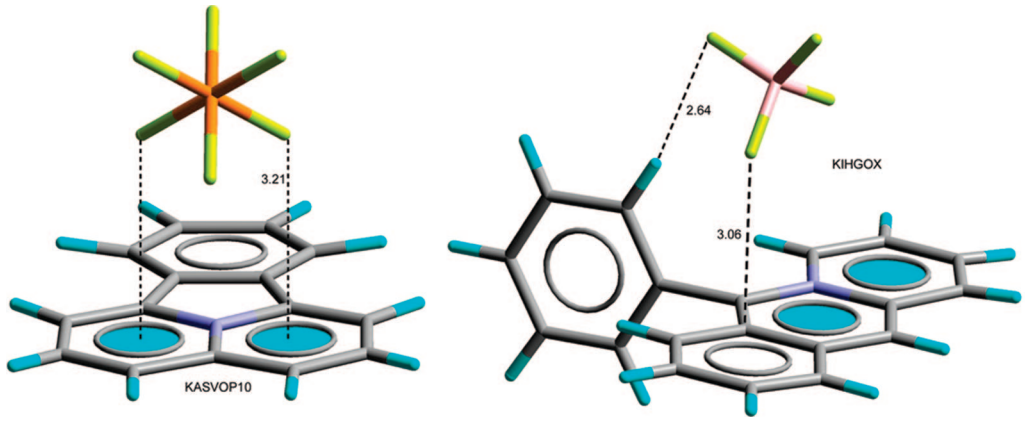



***Figure 6.*** Partial views of the X-ray structures corresponding to **10c**−azoniafluoranthene hexafluorophosphate (KASVOP10, left) and 1-phenylbenzo(**b**)quinolizinium tetrafluoroborate (KIHGOX).

CP connects the fluorine atom with the center of the ring. Second is the interaction of a third fluorine atom of the anion with several atoms of the ring via three bond and two ring CPs in $\mathbf{4c_1}$ and $\mathbf{5c_1}$ and via one bond CP in $\mathbf{4c_2}$ and $\mathbf{5c_2}$. These CPs are located outside the molecular projection in the on-top representation shown in Figure 5. Finally, the third interaction involves the boron atom of the anion, which is connected with **1** through three ring CPs and one cage CP. Quantitative values for $\rho(r)$ and $\nabla^2\rho(r)$ at the CPs give hints on the character and strength of the interaction. These values are summarized in Table 2 for selected CPs that characterize the noncovalent interaction. In all complexes, the value of the Laplacian at the (3, −1) CPs is positive, indicating a depletion of the electron density, as is common in closed-shell interactions. From the data summarized in Table 2, several considerations can be inferred. First, the values of $\rho(r)$ and $\nabla^2\rho(r)$ at the bond CPs in chloride complexes **2a**−**2c** are higher than the ones for bromide complexes **3a**−**3c**, in agreement with the differences in the interaction energies and equilibrium distances. Second, the same is applicable to hydrogen bonding complexes **4b** and **5b** and $\pi$-complexes **4c** and **5c**. For latter complexes ($\mathbf{4c_{1-2}}$ and $\mathbf{5c_{1-2}}$), we have used the cage CP because previous studies have demonstrated that the value of the electron charge density at the cage CP can be used as a measure of the bond order in anion−$\pi$ complexes,[5,7 41] and cation−$\pi$ interactions.[42] The absolute values of $\rho(r)$ and $\nabla^2\rho(r)$ of anion−$\pi$ complexes **2c**−**5c** are greater than the ones previously reported in neutral systems like hexafluorobenzene. Moreover, taking into account that the equilibrium distances observed in the anion−$\pi$ complexes **2c**−**5c** are shorter than the previously reported for standard anion−$\pi$ complexes, it could be assumed that these complexes contain some degree of covalent character. A more likely explanation is that the interaction of anions with the charged arene **1** involves a strong electrostatic attraction that shortens the equilibrium distances of the complexes. In fact, the values of equilibrium distances and charge density at the CPs present in Tables 1 and 2 are in agreement with the ones obtained for the interaction of anions with the tropylium cation.[23]

**D. CSD Analysis.** To obtain experimental evidence of the anion-$\sigma/\pi$ dual binding affinity of **1**, we performed a search in the Cambridge Structural Database (CSD).[43] Crystal structures are so rich in geometrical information and often reveal effects that have not been noticed by the original authors. The utility of crystallography and the CSD in analyzing geometrical parameters and noncovalent interactions is clearly established.[44] In exploring the CSD for derivatives of quinolizinylium, we have found 23 structures that have been analyzed to determine if the counterion is establishing hydrogen-bonding or anion−$\pi$ interactions. In Table 3, we summarize the CSD reference codes, the counterion, and the type of noncovalent interaction. It can be observed that spherical anions give in all cases $\sigma$ interactions, and conversely, they do not participate in anion−$\pi$ interactions. In contrast, all polyatomic anions participate in anion−$\pi$ interactions, apart from entry 8. In this structure, the $\pi$ system is not available for anion−$\pi$ bonding because it establishes an intermolecular $\pi$−$\pi$ stacking that controls the crystal packing. The polyatomic anions, in addition to their participation in $\pi$-interactions, are also involved in a variety of $\sigma$-interactions with other quinolizinylium moieties and a combination of both interactions controls the crystal packing. Two selected examples retrieved from the CSD (codes KASVOP10[45] and KIHGOX[46]) are shown in Figure 6, in which the anion−$\pi$ interaction is evident and plays a prominent role in the crystal packing. It is remarkable the agreement between the solid state geometry of KASPOV10 and the optimized complex $\mathbf{5c_2}$. In the solid state, the equilibrium distance is larger because the anion is participating simultaneously in a variety of hydrogen bonding interactions with neighboring quinolizinylium moieties.

## IV. Concluding Remarks

In summary, we have studied the dual $\sigma/\pi$ binding ability of the quinolizinylium cation (**1**), by means of MP2 ab initio calculations, MIPp, and IPE energy maps, and the AIM theory. In addition, we have analyzed the X-ray crystal structures present in the CSD. We have theoretically demonstrated that hydrogen-bonding interactions are more favorable in the complexes of **1** with monatomic anions. In contrast, $\pi$-interactions are more favorable in the complexes of **1** with tetrahedral $BF_4^-$ and octahedral $PF_6^-$ anions. These results are in agreement with experimental data obtained from

the CSD. We have demonstrated the utility of MIPp/IPE maps as predictive tools, and we have described the interactions by using the distribution of critical points that emerge upon complexation.

**Acknowledgment.** We thank the DGICYT of Spain (projects CTQ2005-08989-01 and CTQ2005-08989-02) for financial support. We thank the CESCA for computational facilities. We thank the Govern Balear (project PROGECIB-33A) for financial support.

**Supporting Information Available:** Cartesian coordinates of RI-MP2/6−31++G** optimized structures **1**−**5**. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.

(2) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210.

(3) Mascal, M.; Armstrong, A.; Bartberger, M. *J. Am. Chem. Soc.* **2002**, *124*, 6274.

(4) Alkorta, I.; Rozas, I.; Elguero, J. *J. Am. Chem. Soc.* **2002**, *124*, 8593.

(5) Quiñonero, D.; Garau, C.; Rotger, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Angew. Chem., Int. Ed.* **2002**, *41*, 3389.

(6) Hiraoka, K.; Mizuse, S.; Yamabe, S. *J. Phys. Chem.* **1987**, *91*, 5294.

(7) Schneider, H. J.; Yatsimirski, A. *Principles and Methods in Supramolecular Chemistry*; John Whiley & Sons Ltd: Chichester, U.K., 2000; pp 93−95.

(8) (a) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *ChemPhysChem* **2003**, *4*, 1344. (b) Escudero, D.; Frontera, A.; Quiñonero, D.; Costa, A.; Ballester, P.; Deyà, P. M. *J. Chem. Theory Comput.* **2007**, *36*, 2098.

(9) Demeshko, S.; Dechert, S.; Meyer, F. *J. Am. Chem. Soc.* **2004**, *126*, 4508.

(10) Schottel, B. L.; Bacsa, J.; Dunbar, K. R. *Chem. Commun.* **2005**, 46.

(11) Rosokha, Y. S.; Lindeman, S. V.; Rosokha, S. V.; Kochi, J. K. *Angew. Chem., Int. Ed.* **2004**, *43*, 4650.

(12) de Hoog, P.; Gamez, P.; Mutikainen, I.; Turpeinen, U.; Reedijk, J. *Angew. Chem., Int. Ed.* **2004**, *43*, 5815.

(13) (a) Frontera, A.; Saczewski, F.; Gdaniec, M.; Dziemidowicz-Borys, E.; Kurland, A.; Deyà, P. M.; Quiñonero, D.; Garau, C. *Chem.—Eur. J.* **2005**, *11*, 6560. (b) Gil-Ramirez, G.; Benet-Buchholz, J.; Escudero-Adan, E. C.; Ballester, P. *J. Am. Chem. Soc.* **2007**, *129*, 3820.

(14) García-Raso, A.; Albertí, F. M.; Fiol, J. J.; Tasada, A.; Barceló-Oliver, M.; Molins, E.; Escudero, D.; Frontera, A.; Quiñonero, D.; Deyà, P. M. *Inorg. Chem.* **2007**, *46*, 10724.

(15) Berryman, O. B.; Bryantsev, V. S.; Stay, D. P.; Johnson, D. W.; Hay, B. P. *J. Am. Chem. Soc.* **2007**, *129*, 48.

(16) (a) Gamez, P.; Mooibroek, T. J.; Teat, S. J.; Reedijk, J. *Acc. Chem. Res.* **2007**, *40*, 435. (b) Schottel, B. L.; Chifoides, H. T.; Dunbar, K. R. *Chem. Soc. Rev.* **2008**, *37*, 68. (c) Hay, B. P.; Bryantsev, V. S. *Chem. Commun.* **2008**, 2417.

(17) (a) Tabcharani, J. A.; Chang, X.-B.; Riordan, J. R.; Hanrahan, J. W. *Nature* **1991**, *352*, 628. (b) Quinton, P. M. *Physiol. Rev.* **1999**, *79*, S3.

(18) Welsh, M. J.; Smith, A. E. *Cell* **1993**, *73*, 1251.

(19) Marivingt-Mounir, C.; Norez, C.; Derand, R.; Bulteau-Pignoux, L.; Nguyen-Huy, D.; Viossat, B.; Morgant, G.; Becq, F.; Vierfond, J.-M.; Mettey, Y. *J. Med. Chem.* **2004**, *47*, 962.

(20) Alkorta, I.; Blanco, F.; Elguero, J. *J. Phys. Chem. A* **2008**, *112*, 1817.

(21) (a) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2004**, *108*, 9423. (b) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *Chem. Phys. Lett.* **2004**, *392*, 85.

(22) (a) Mascal, M.; Yakovlev, I.; Nikitin, E. B.; Fettinger, J. C. *Angew. Chem., Int. Ed.* **2007**, *46*, 8782. (b) Mascal, M. *Angew. Chem., Int. Ed.* **2006**, *45*, 2890. (c) Berryman, O. B.; Hof, F.; Hynes, M. J.; Johnson, D. W. *Chem. Commun.* **2006**, 506.

(23) Quiñonero, D.; Frontera, A.; Escudero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *ChemPhysChem* **2007**, *8*, 1182.

(24) Ahlrichs, R.; Bär, M.; Hacer, M.; Horn, H.; Kömel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(25) Feyereisen, M. W.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359.

(26) Vahtras, O.; Almlof, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514.

(27) Frontera, A.; Quiñonero, D.; Garau, C.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2005**, *109*, 4632.

(28) Quiñonero, D.; Garau, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2006**, *110*, 5144.

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

(30) Luque, F. J.; Orozco, M. *J. Comput. Chem.* **1998**, *19*, 866.

(31) Scrocco, E.; Tomasi, J. *Top. Curr. Chem.* **1973**, *42*, 95.

(32) Francl, M. M. *J. Phys. Chem.* **1985**, *89*, 428.

(33) Luque, F. J.; Orozco, M. *MOPETE-98 Computer Program*; Universitat de Barcelona: Barcelona, Spain, 1998.

(34) (a) Clark, M.; Cramer III, R. D.; Opdenbosch, N. *J. Comput. Chem.* **1989**, *10*, 982. (b) Ujaque, G.; Maseras, F.; Eisenstein, O. *Theor. Chem. Acc.* **1997**, *96*, 146.

(35) (a) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893. (b) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon: Oxford, U.K., 1990.

(36) http://www.AIM 2000.de (accessed Sep 17, 2008).

(37) Sigfridson, E.; Ryde, U. *J. Comput. Chem.* **1998**, *19*, 377.

(38) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(39) (a) Emmeluth, C.; Poad, B. L. J.; Thompson, C. D.; Bieske, E. J. *J. Phys. Chem. A* **2007**, *111*, 7322. (b) Thompson, C. D.; Poad, B. L. J.; Emmeluth, C.; Bieske, E. J. *Chem. Phys. Lett.* **2006**, *428*, 18. (c) Loh, Z. M.; Wilson, R. L.; Wild, D. A.; Bieske, E. J.; Zehnacker, A. *J. Chem. Phys.* **2003**, *119*, 9559. (d) Schneider, H.; Vogelhuber, K. M.; Schinle, F.; Weber, J. M. *J. Am. Chem. Soc.* **2007**, *129*, 13022.

(40) Bader, R. F. W. *J. Phys. Chem. A* **1998**, *102*, 7314.

(41) Quiñonero, D.; Garau, C.; Frontera, A.; Ballester, P.; Costa, A.; Deya, P. M. *Chem. Phys. Lett.* **2002**, *359*, 486.

(42) Cubero, E.; Orozco, M.; Luque, F. J. *J. Phys. Chem. A* **1999**, *103*, 315.

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

(44) Nangia, A.; Biradha, K.; Desiraju, G. R. *J. Chem. Soc., Perkin Trans.* **2 1996**, 943–953.

(45) Fourmigue, M.; Boubekeur, K.; Batail, P.; Bechgaard, K. *Angew. Chem., Int. Ed. Engl.* **1989**, *28*, 588.

(46) Maassarani, F.; Pfeffer, M.; Le Borgne, G. *Organometallics* **1990**, *9*, 3003.

(47) Elix, J. A.; Sterns, M.; Wilson, W. S.; Warrener, R. N. *J. Chem. Soc. D* **1971**, 426.

(48) Ihmels, H.; Leusser, D.; Pfeiffer, M.; Stalke, D. *J. Org. Chem.* **1999**, *64*, 5715.

(49) Florencio, F.; Smith-Verdier, P.; Garcia-Blanco, S. *Z. Kristallogr.* **1984**, *167*, 29.

(50) Granzhan, A.; Ihmels, H.; Mikhlina, K.; Deiseroth, H.-J.; Mikus, H. *Eur. J. Org. Chem.* **2005**, 4098.

(51) Sato, K.; Arai, S.; Yamagishi, T.; Tanase, T. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **2001**, *57*, 174.

(52) Granzhan, A.; Bats, J. W.; Ihmels, H. *Synthesis* **2006**, 1549.

(53) Ihmels, H.; Mohrschladt, C. J.; Schmitt, A.; Bressanini, M.; Leusser, D.; Stalke, D. *Eur. J. Org. Chem.* **2002**, 2624.

(54) Dai, W.; Petersen, J. L.; Wang, K. K. *J. Org. Chem.* **2005**, *70*, 6647.

(55) Ihmels, H.; Faulhaber, K.; Wissel, K.; Viola, G.; Vedaldi, D. *Org. Biomol. Chem.* **2003**, *1*, 2999.

(56) Yoshino, H.; Koike, K.; Nikaido, T. *Heterocycles* **1999**, *51*, 281.

(57) Sato, K.; Arai, S.; Yamagishi, T.; Tanase, T. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **2003**, *59*, o162.

CT800332Y

JCTC Journal of Chemical Theory and Computation

**CHARMM Additive All-Atom Force Field for Acyclic Carbohydrates and Inositol.** [*J. Chem. Theory Comput.* *4,* 765–778 (2008)]. By Ganesh Kamath, Olgun Guvench, and Alexander D. MacKerell, Jr.*.

Pages 765−778. The article was retracted on October 27, 2008 due to violations of the ACS Ethical Guidelines.

CT800423M

10.1021/ct800423m
Published on Web 11/11/2008